Molecular Breeding

Marina Ćeran and Aleksandra Radanović

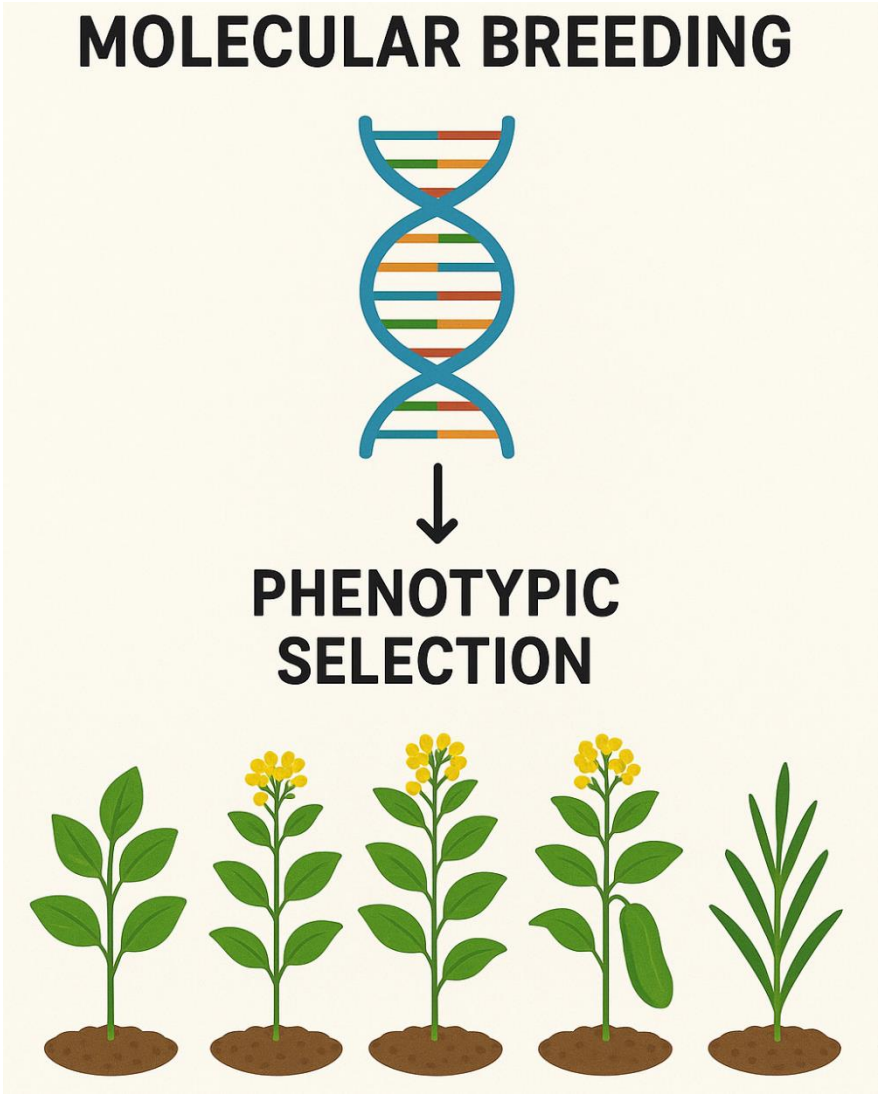
OVERVIEW

I Introduction to molecular breeding

II Application of molecular markers in crop breeding

III The Omics Revolution in Crop Breeding

IV Genome editing



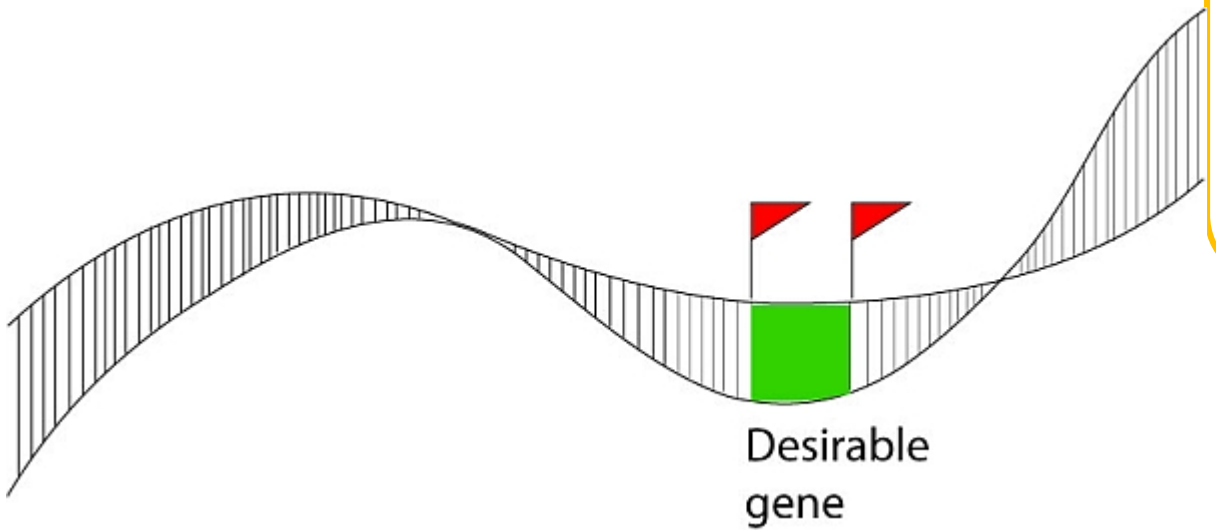
- improve selection efficiency
- reducing field trials and cost
- speed-up breeding process
- independent of environmental conditions

Molecular breeding may be defined in a broad-sense as the use of genetic manipulation performed at the level of DNA to improve traits of interest in plants and animals, and it may also include genetic engineering or gene manipulation, molecular marker-assisted selection, and genomic selection.

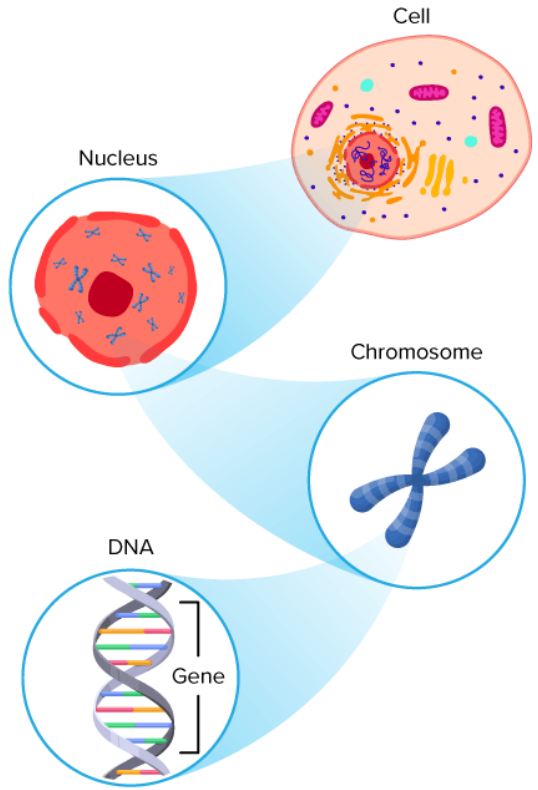
Jiang, G.L., 2013. Molecular markers and marker-assisted breeding. *Jiang, G.-L. Plant Breeding from Laboratories to Fields*, pp.45-85.

Molecular markers

-fragment of DNA at a certain location within the genome
-polymorphism in DNA that can be easily tracked and quantified
-insertions, deletions, translocations, duplications and point mutations
-may correlate with a particular gene or trait of interest

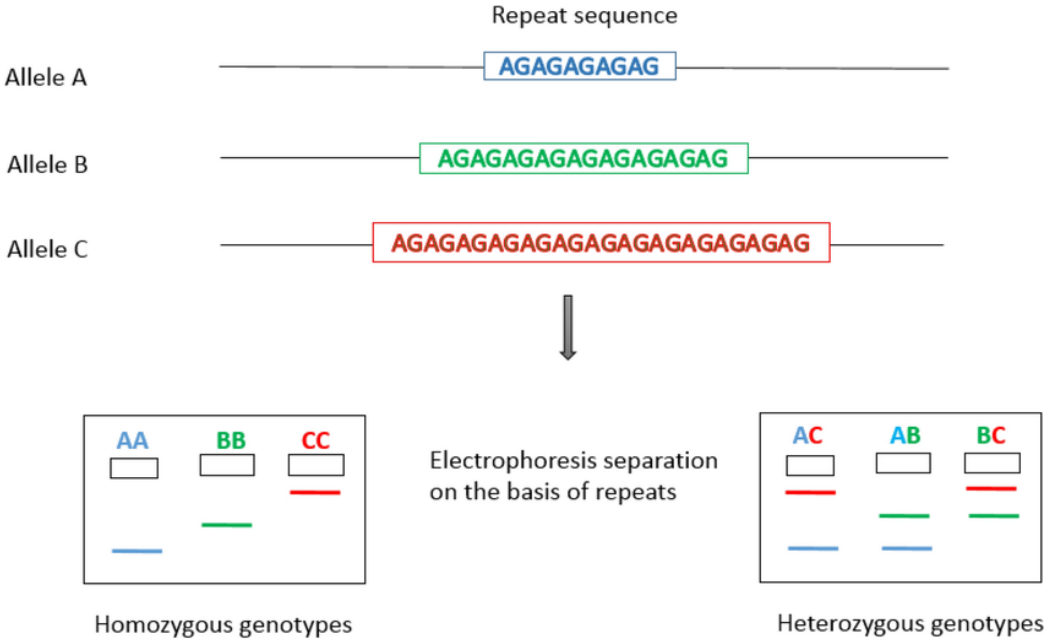


- highly polymorphic
- abundant throughout the entire genome
- not confounded by environmental, pleiotropic and epistatic effects
- stable, detectable at any time during plant development and from any organ or tissue



Simple sequence repeats (SSR)

- consists of short **DNA motifs** (1–6 base pairs) that are **repeated in a tandem fashion**, such as ATATAT or ACGTACGT.
- the number of these repeats varies significantly between individuals, making **SSRs highly polymorphic** and ideal for genetic studies
- Because they are **co-dominant**, both alleles in a heterozygous individual can be expressed and detected, allowing for precise identification of genotypes
- SSRs are developed from either genomic DNA (gSSRs) or expressed sequence tags (EST-SSRs), with EST-SSRs being easier to develop and more transferable to related species



Task:
•Identify the SSR motif (the repeating unit).
•Count the number of repeats in each sequence.
•Design a pair of forward and reverse PCR primers for each sequence.
•Predict the length of the PCR product for each sequence.

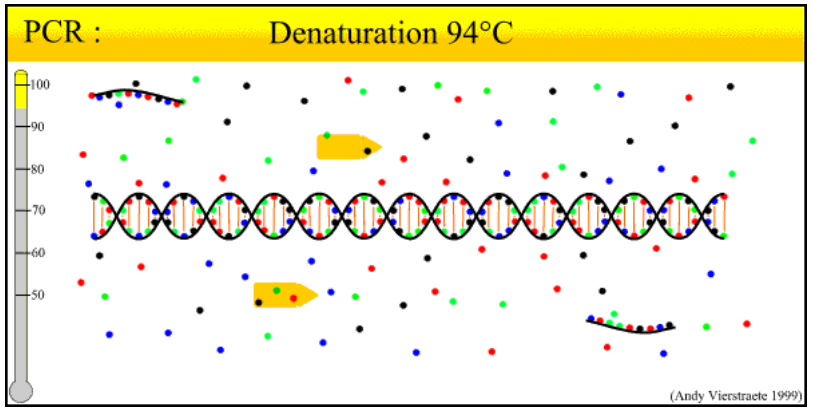
Sequence 1: "Allele A"

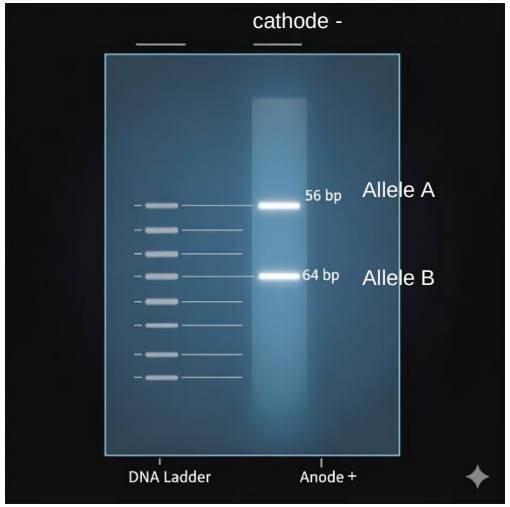
GCTAGCTAATTCGTACGGGGGAGAGAGAGAGAGAGAT
ACATACCGCTAGGCATTCG

Sequence 2: "Allele B"

GCTAGCTAATTCGTACGGGGGAGAGAGAGAGAGAGAGAGAGAGAGCCA
TACCGCTAGGCATTCG

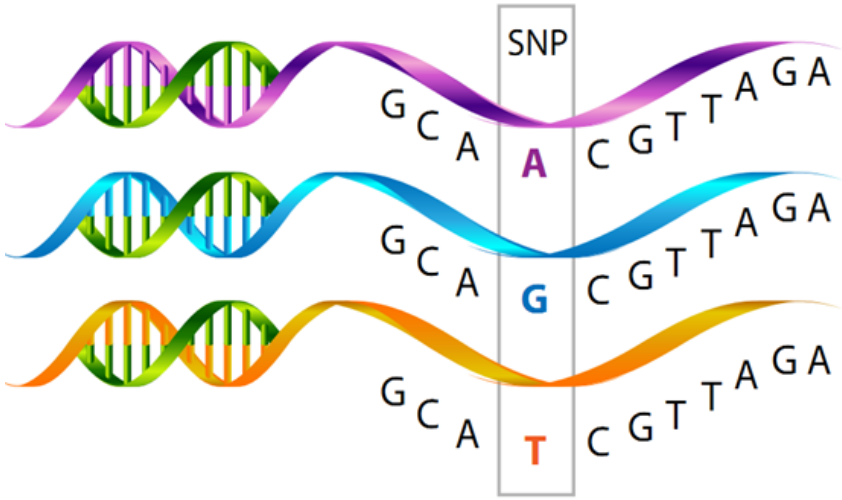
SSR motif: GA
Number of repeats: 8 and 12
Forward primer – GCTAGCTAATTCGTACGG
Reverse primer - reverse compliment of CATACCGCTAGGCATTCG - CGAATGCCTAGCGGTATG

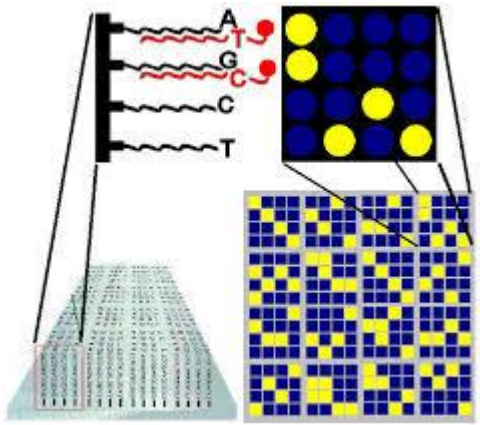




Single nucleotide polymorpism (SNP)

- single base-pair differences in DNA
- high abundance in the genome
- easy detection
- widely dispersed throughout genomes





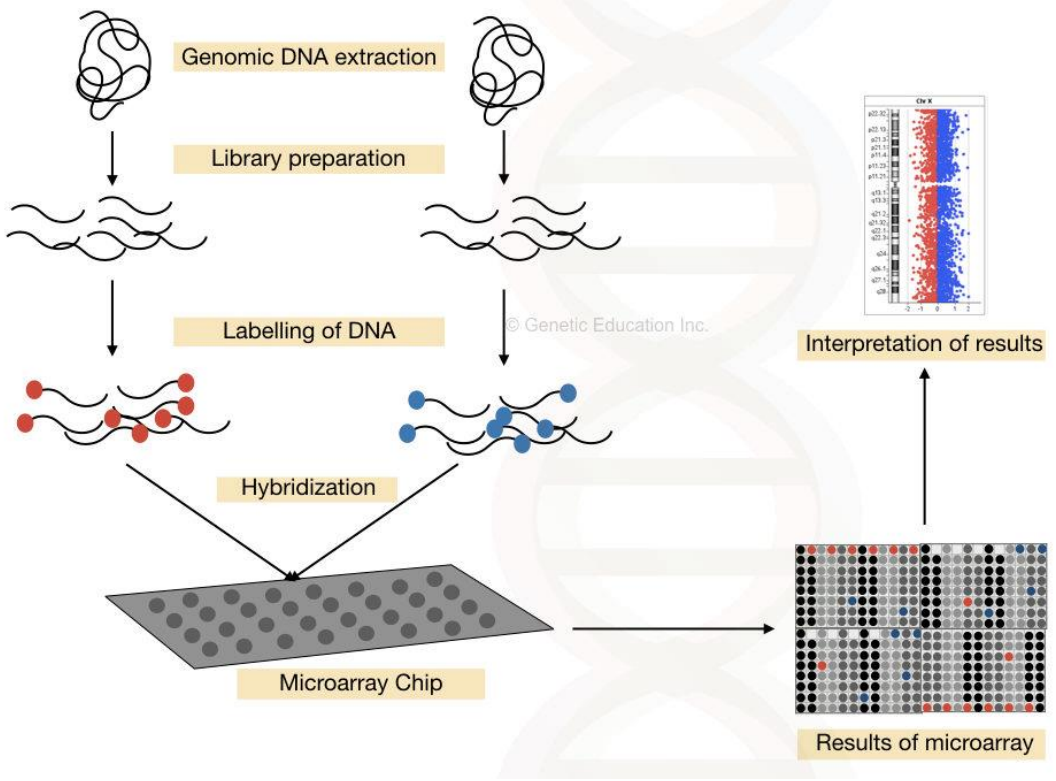
- Hybridization-based methods (SNP microarrays)
- Enzyme-based methods (PCR-based methods)
- Post-amplification methods based on physical properties of DNA (HRM)
- DNA Sequencing

Hybridization-based methods (SNP microarrays)

•Sample Preparation: A plant's genomic DNA is extracted, fragmented, and labeled with a fluorescent tag.

•Hybridization: The labeled DNA fragments are washed over the microarray. A fragment will bind (or hybridize) only to a probe on the chip that has a perfectly matching DNA sequence.

•Detection: A specialized scanner detects the fluorescent signal. The presence and intensity of a signal at a particular spot on the chip indicate that the DNA from the sample hybridized to that specific probe.

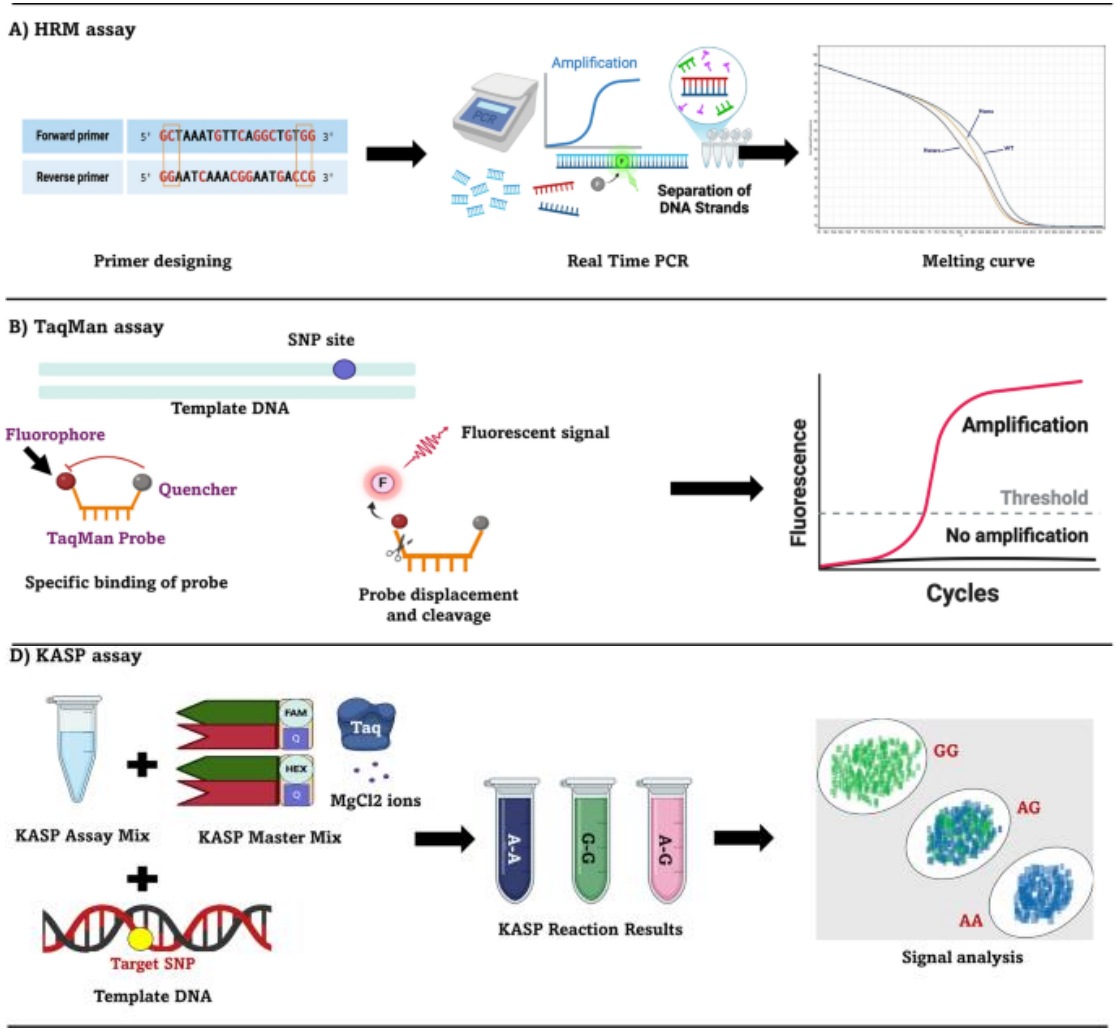


https://geneticeducation.co.in/snp-array-high-throughput-snp-genotyping-technique/

PCR-based methods

A High-resolution melting (HRM) uses intercalating fluorescent dyes for the quantitative analysis of the 'melting curves' of the DNA fragments based on how DNA dissociation takes place from double-stranded DNA (dsDNA) to single-stranded DNA (ssDNA) with increasing temperature.

B TaqMan based assay is based on hybridization of allele-specific probes to the target SNP site. A fluorescent signal is produced when the probes are displaced and cleaved leading to separation of the fluorophore from the quencher due to the 5′ -nuclease activity of the Taq polymerase.



D Kompetitive Allele-Specific PCR (KASP) consists of two allele-specific primers with a unique tail sequence, one reverse primer, two fluorescently labelled oligonucleotides with FAM and HEX which interact with the tail sequence of the allele-specific primers and two oligonucleotides with quenchers at the 3′ ends. Binding between allele-specific primer to its complementary region of the target SNP takes place. Bi-allelic discrimination is achieved through competitive binding and differential of heterozygous and homozygous allele based on single or mixed fluorescence

https://link.springer.com/article/10.1007/s44372-025-00262-9

Next-generation sequencing (NGS) technology

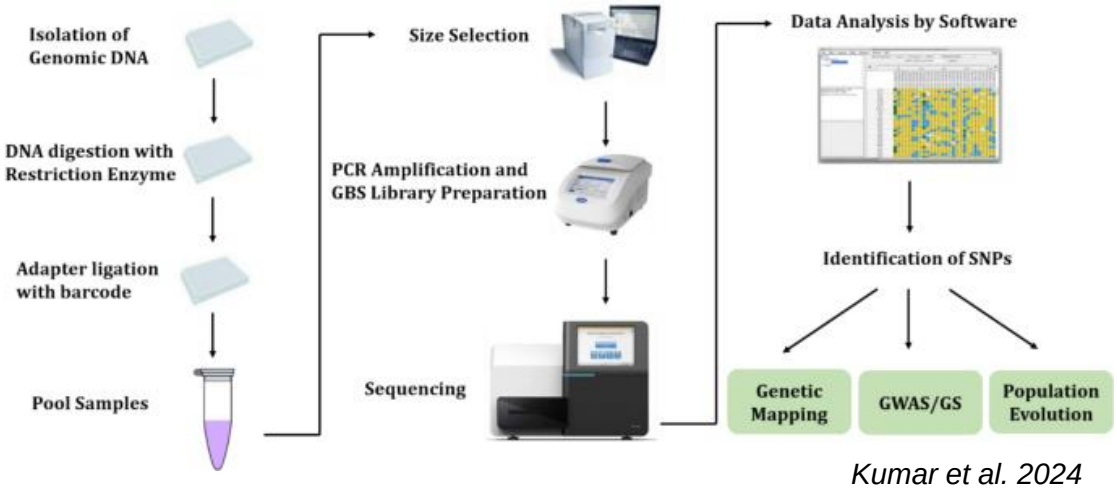
DNA Sequencing

•**Sequencing** is the most comprehensive method for SNP detection. It determines the exact order of nucleotides (A, T, C, G) in a DNA molecule.

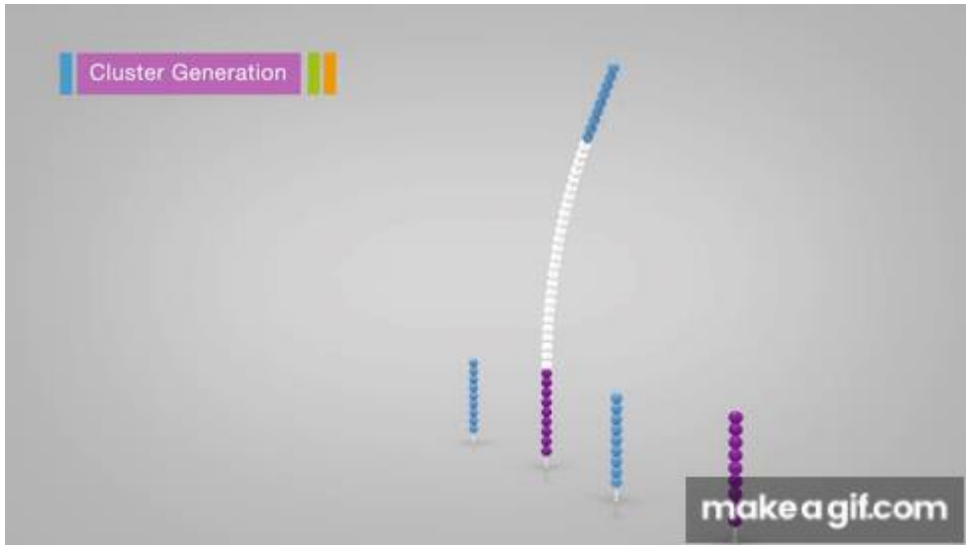
•Instead of just looking for a few known SNPs, sequencing reads a DNA segment and can find **all** of the SNPs within that region. This is called **SNP discovery**.

•**Genotyping-by-Sequencing (GBS)** is a common and powerful strategy. It uses restriction enzymes to cut the genome into smaller, manageable fragments. Only these specific fragments are then sequenced, which makes the process much more efficient and cost-effective than sequencing the entire genome.

•**Why use it?** Sequencing allows for the simultaneous discovery and genotyping of thousands to millions of SNPs at once. This generates a massive amount of data, providing high-resolution genetic information for breeding.



Kumar et al. 2024



SSR vs. SNP: A Comparison for MAS

Polymorphism

SSRs are highly polymorphic (many alleles), making them great for fingerprinting and genetic diversity studies. SNPs are biallelic, which makes them less informative per marker but easier to automate on a massive scale

Cost & Technology

SSR detection with gels is cheaper for a small number of markers, but SNPs are more suited for **high-throughput, automated analysis** using platforms like DNA chips or NGS.

Abundance

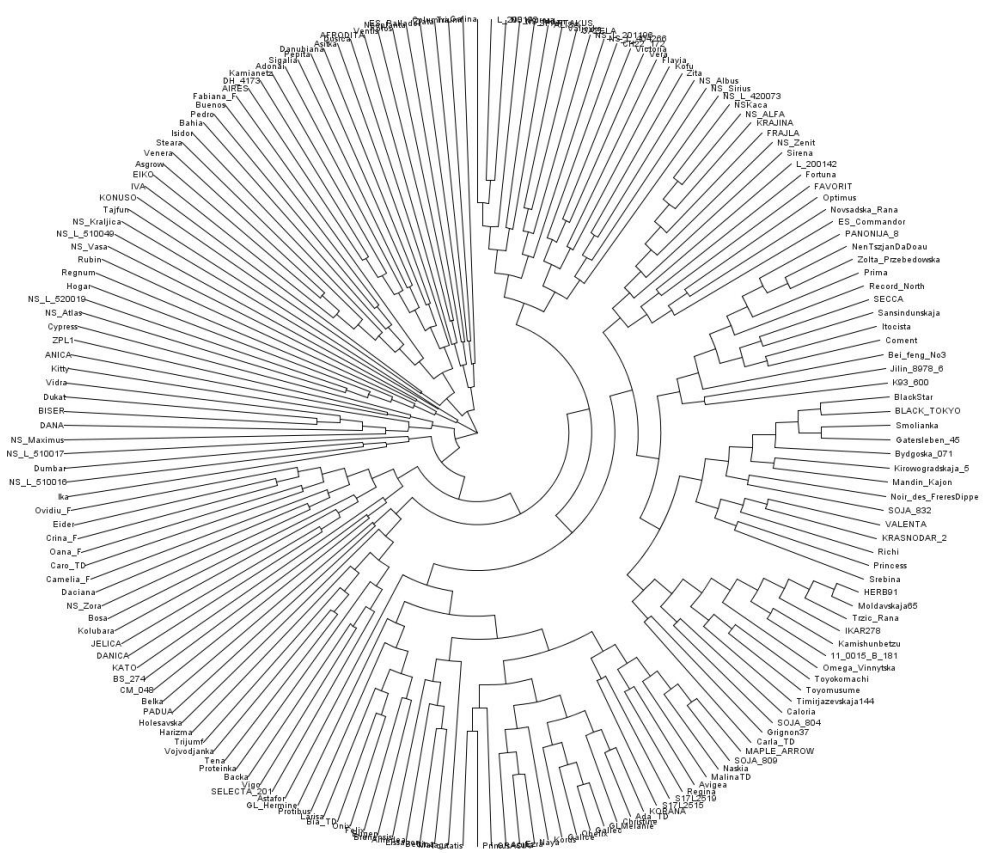
SNPs are far more abundant than SSRs, meaning they provide better coverage of the entire genome.

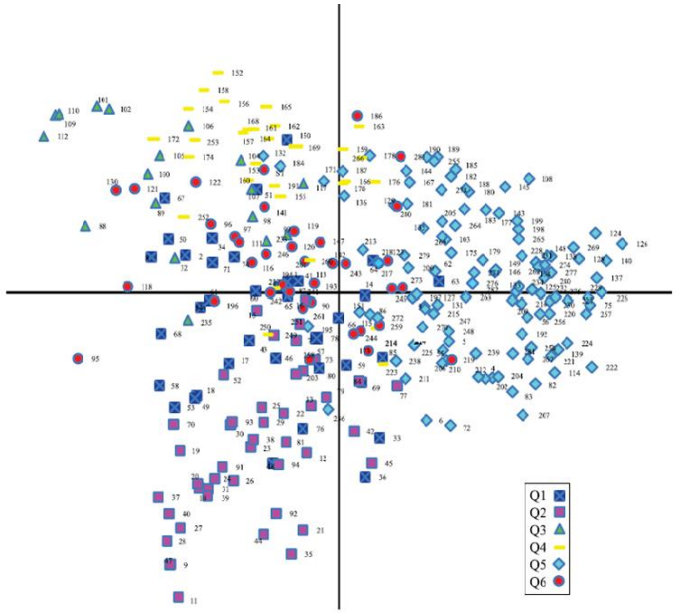
To sum up

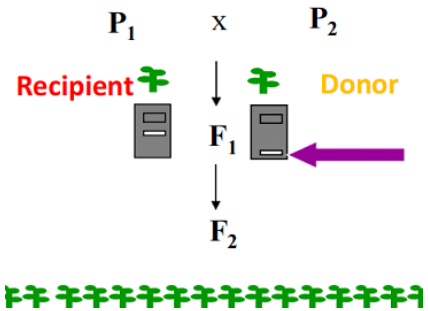
SSRs are great for local studies or fingerprinting, while SNPs have become the dominant marker for large-scale, whole-genome breeding programs.

Applications of molecular markers in crop breeding

1. Detection of genetic diversity
2. DNA fingerprinting
3. Population–genetic studies
4. Genetic mapping
5. Marker-assisted selection
6. Genomic selection
7. Multi-omics data integration
8. Genome editing



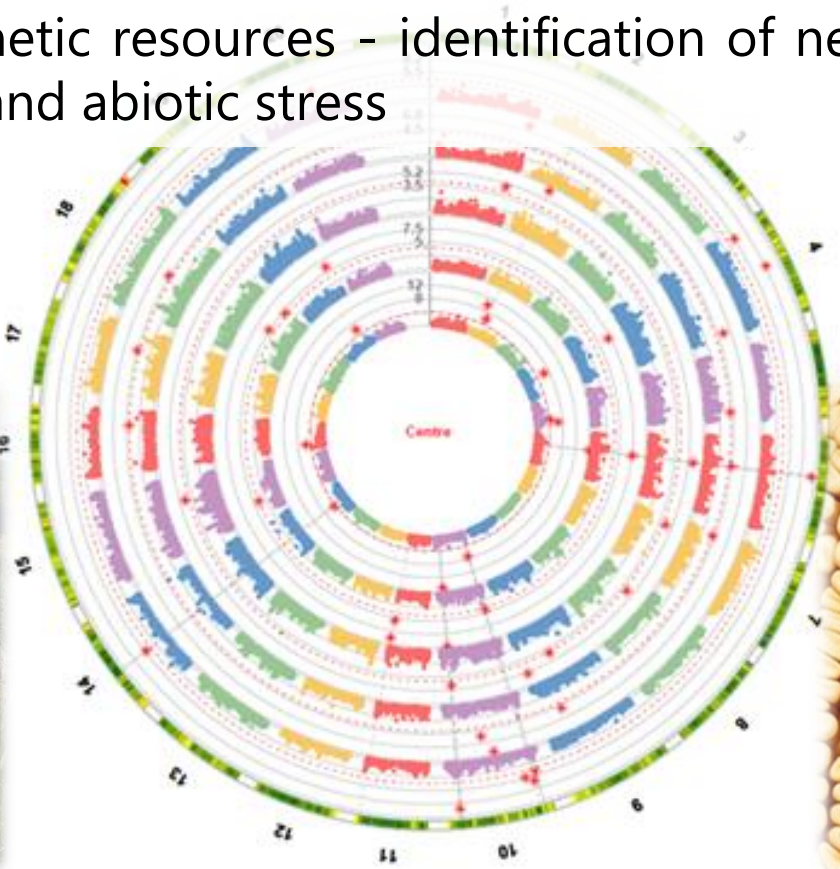




Detection of genetic diversity

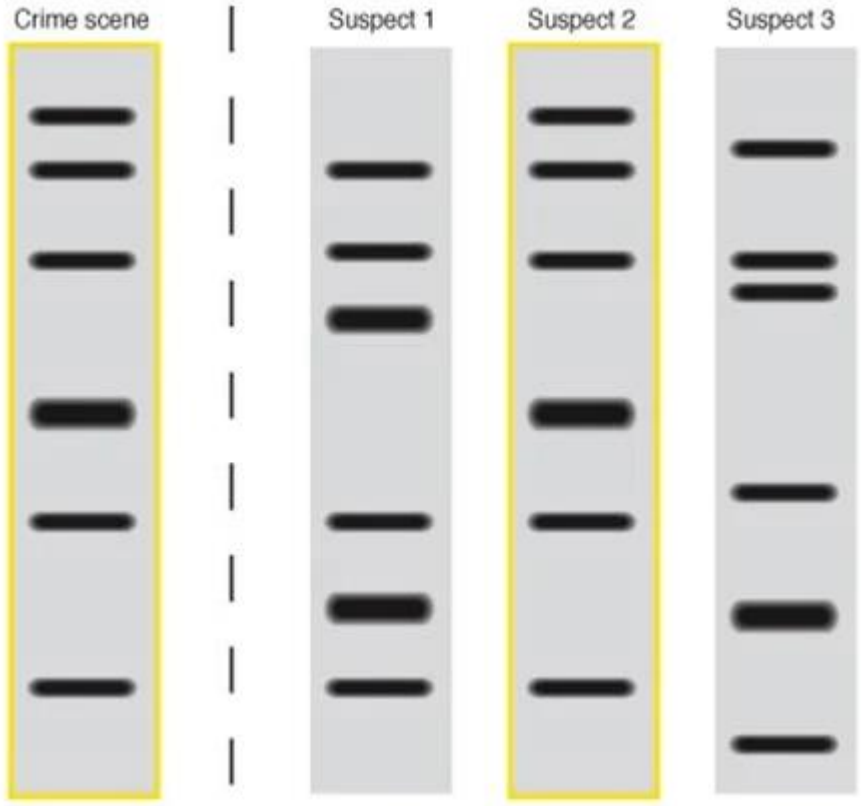
- genetic variability the basic prerequisite for the development of new varieties, esspecially in conditions of changing abiotic and biotic factors

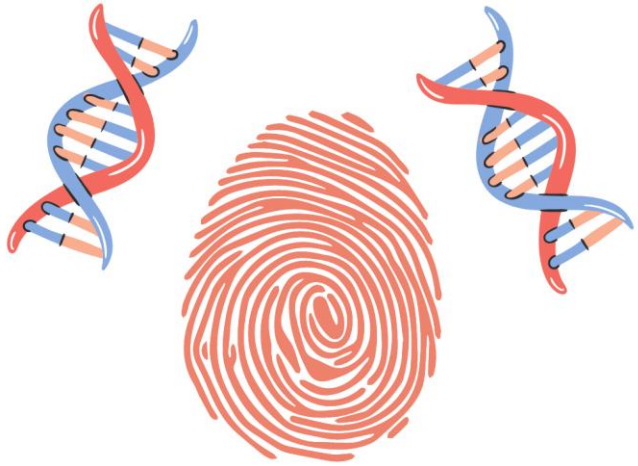
- characterization of genetic resources - identification of new sources of variability or resistance to biotic and abiotic stress

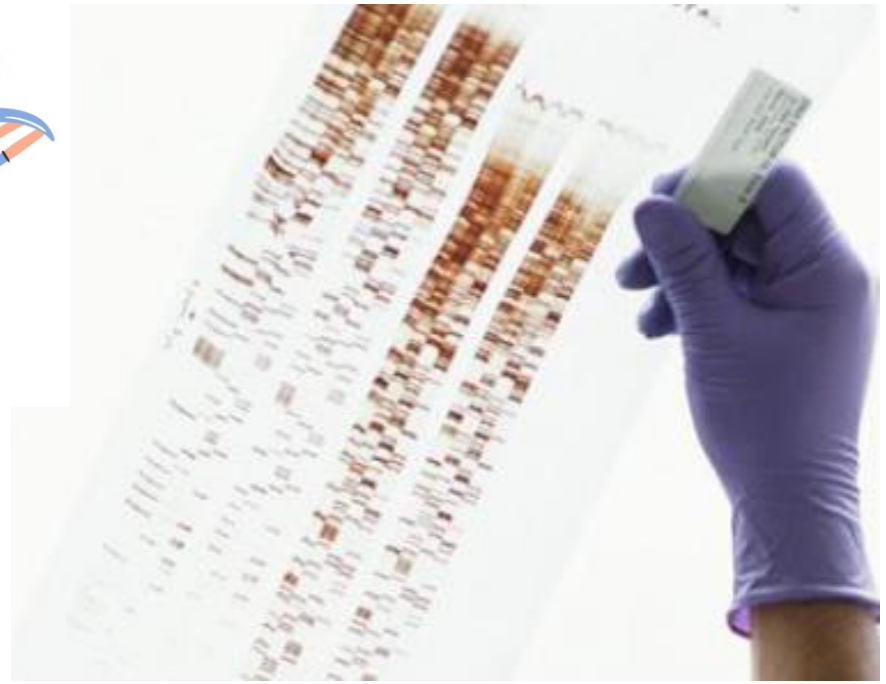


DNA fingerprinting

identification and differentiation of breeding material
based on polymorphism of DNA sequences fully unique to the variety



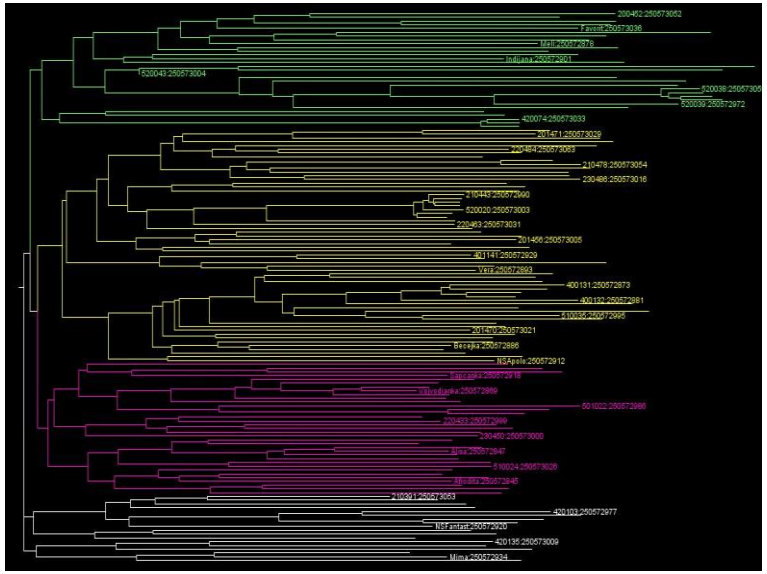


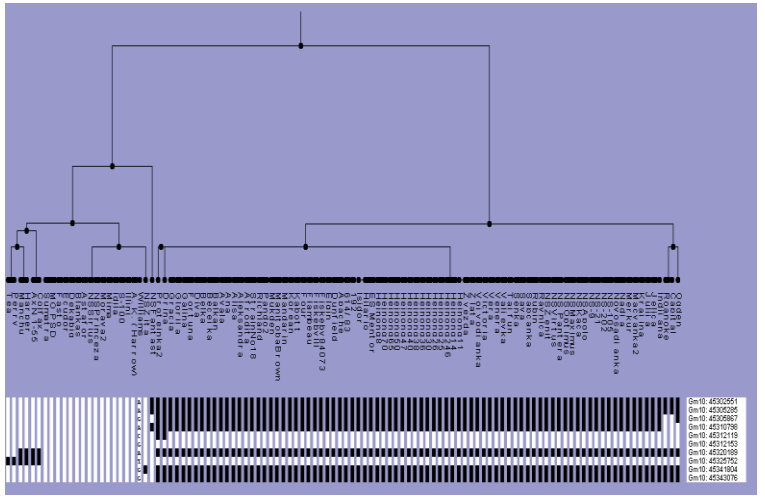


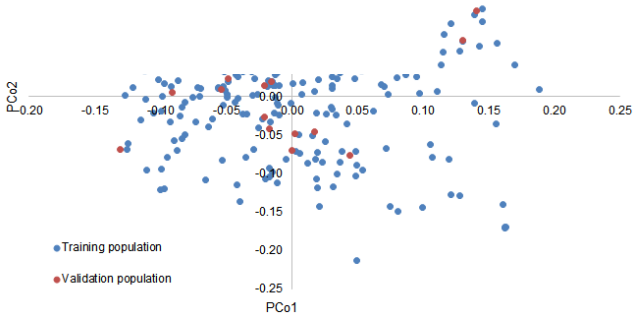
- Plant Variety Protection
- Genetic purity test
- Studying biodiversity
- Conservation of Genetic Resources
- Tracking GMO

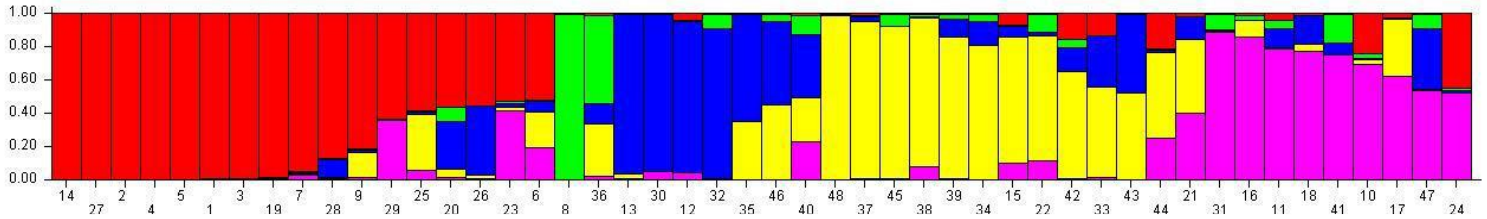
Population–genetic studies

- genetic distance between populations, or breeding material (distance between parents – development of superior genotypes)
- genetic differentation
- genetic structure



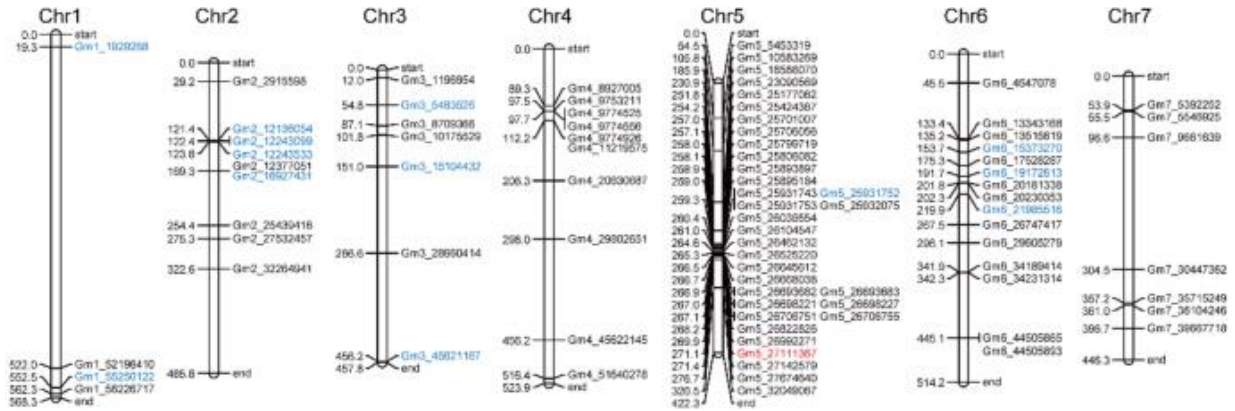




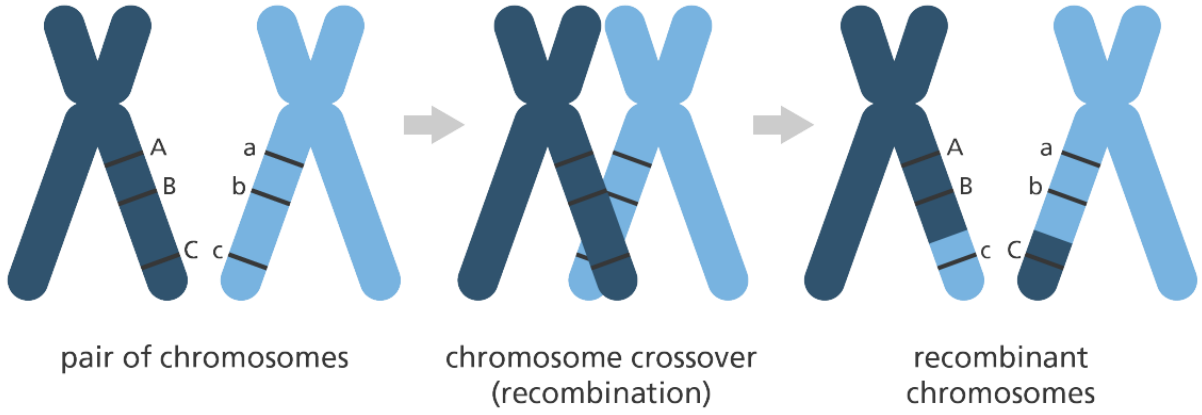


Genetic mapping

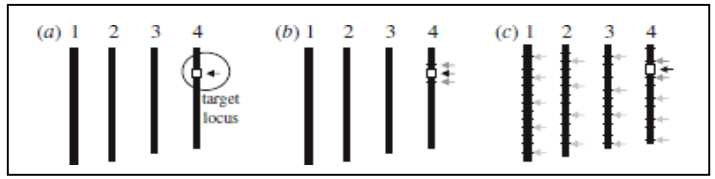
- determination of **distance and linear order** of genes and molecular markers along the chromosome
- distances between adjacent genes proportional to the **frequency of recombination**
- **"marker-trait" associations** - if a particular gene is close to DNA marker on the chromosome, it is more likely that the gene and marker will stay together during the recombination



Li et al. 2019. BMC Genomics 20: 987

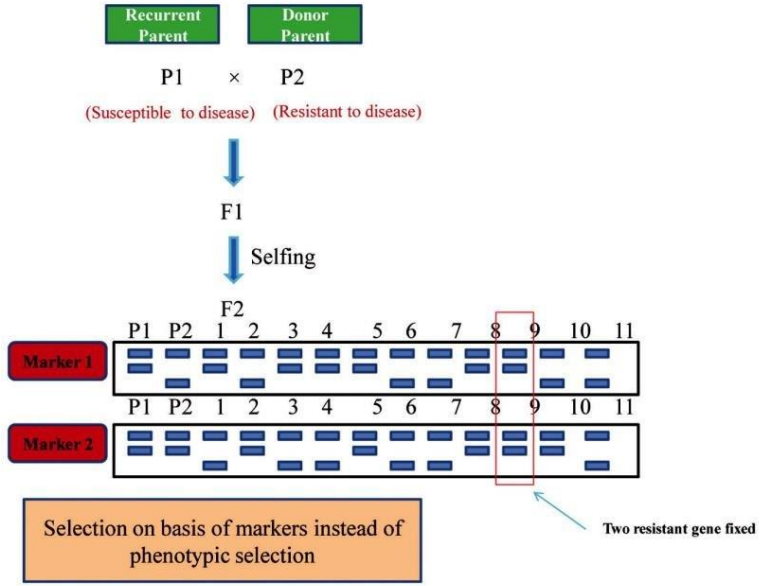


- selection of parents for crossing
- marker-assisted backcrossing (MABC)



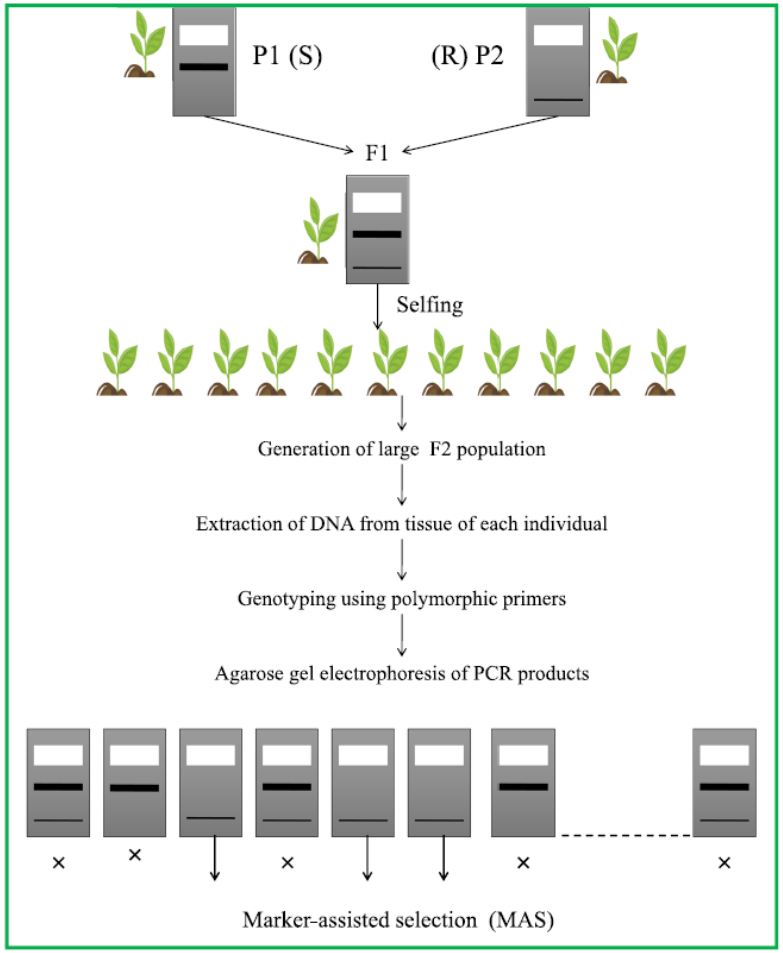
(a) target selection, (b) recombinant selection and (c) background selection (Collard and Mackill 2008)

- Marker-assisted pyramiding (MAGP) - integrating multiple genes or QTLs into a single genotype



Marker-assisted selection (MAS)

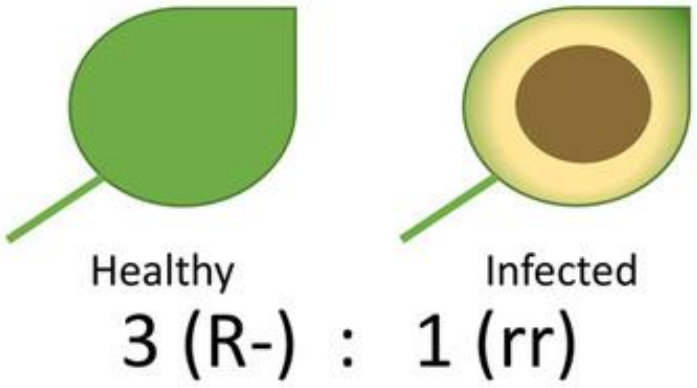
- Precise selection of superior lines early in the breeding
- reducing field trials and cost
- speed-up breeding process

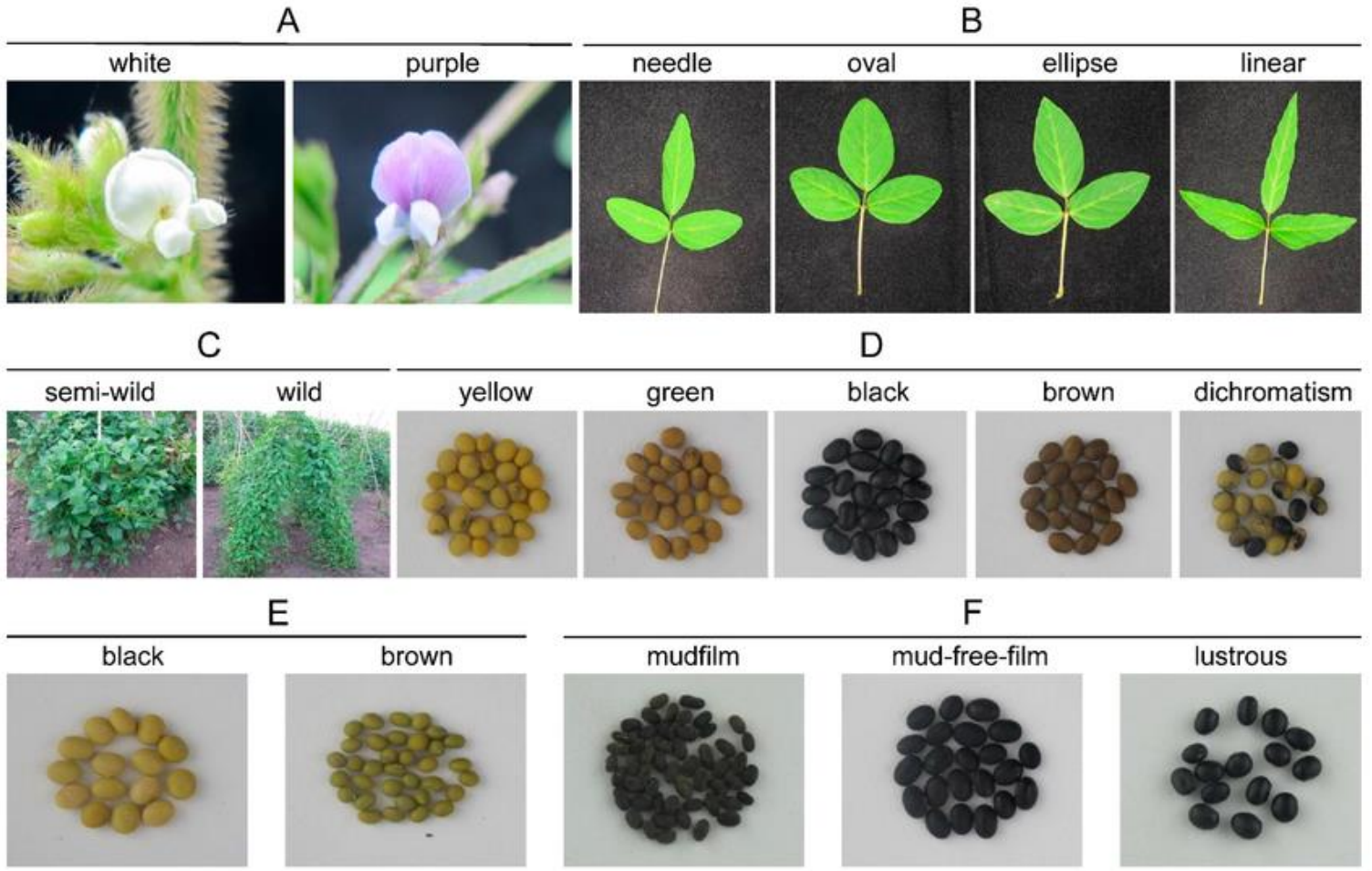


Hasan et al. 2021
Journal of Genetic Engineering and Biotechnology 19:128

Qualitative traits

- single major gene
- discrete values
- simply inherited, not influenced by environmental factors
- segregate at expected Mendelian ratios

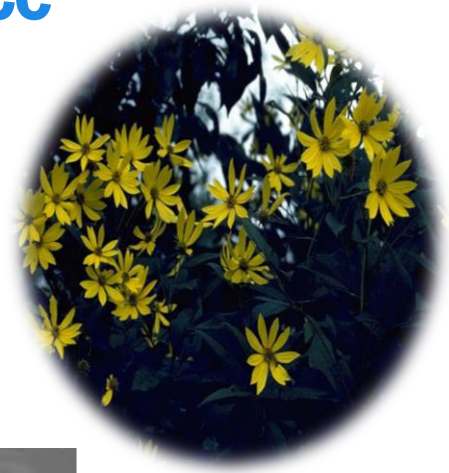




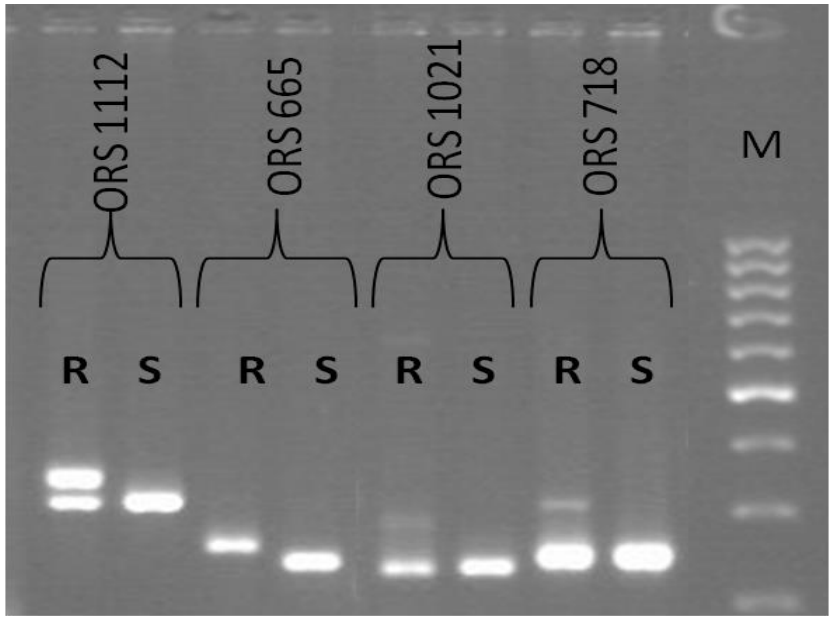
Li et al. 2021, Agronomy

Broomrape resistance

LG3

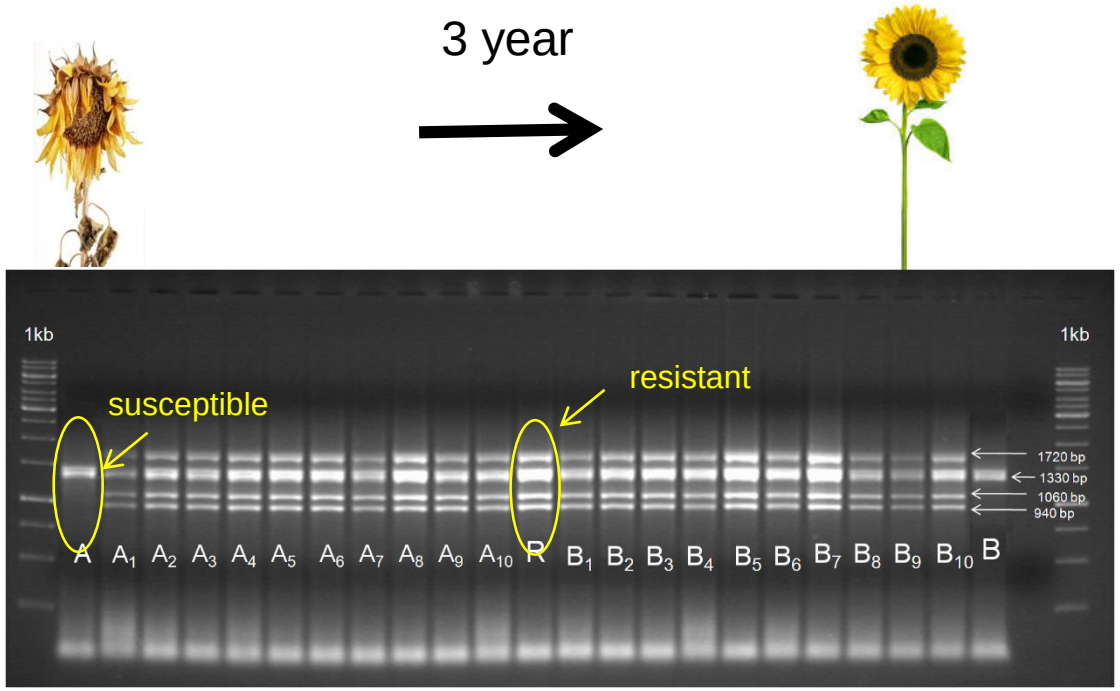


H.divarticatus L.

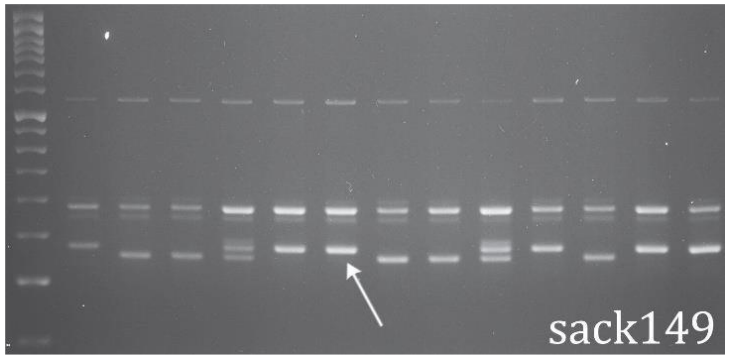


H. argophyllus

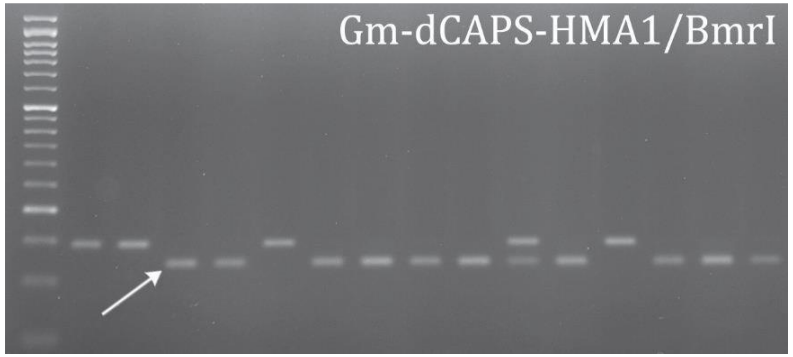
Downy mildew resistance



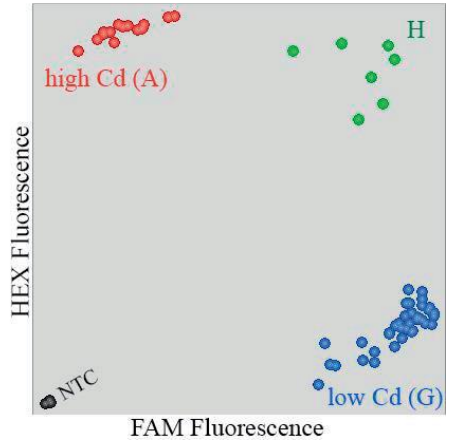
MAS for low Cd accumulation in soybean



Electrophoretic profile SSR-Sack149



Electrophoretic profile obtained with Gm-CAPS-HMA1

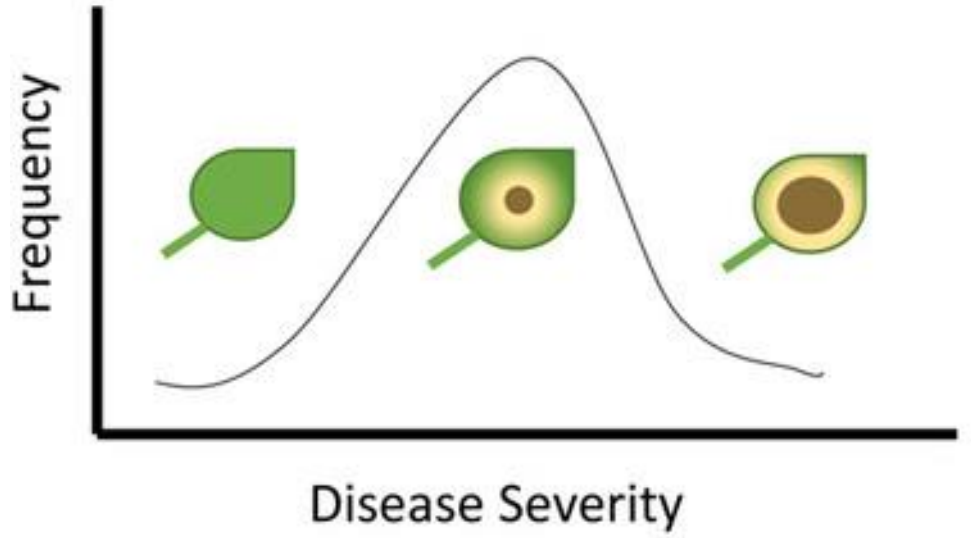


KASP assay for Cda1 alleles differentiation

Ciucă et al. 2023

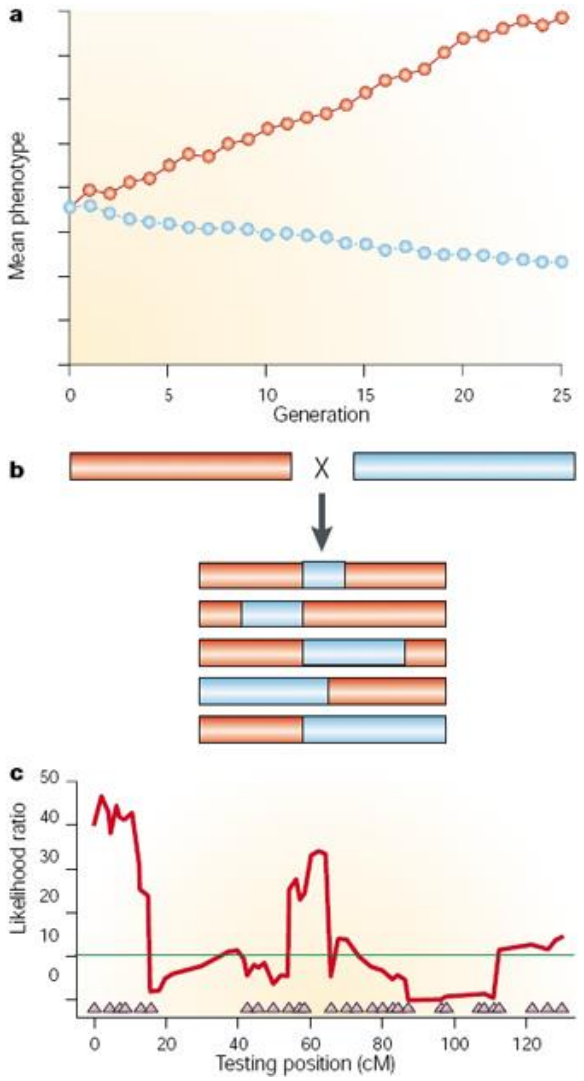
Quantitative Traits

- **Polygenic** (controlled by multiple loci), **influenced by the environment**
- **quantitative trait locus (QTL)** is the genomic location associated with quantitative trait
- each QTL contribute with small effect to the variation of the trait
- phenotypes with **continuous range** and normal distribution
- QTLs affect only a portion of the variability of a trait, phenotypic selection can be rather complicated



Strategies for QTL mapping

Linkage mapping



Miles & Wayne (2008) *Nature Education* 1(1):208

1) creation of **bi-parental** mapping population

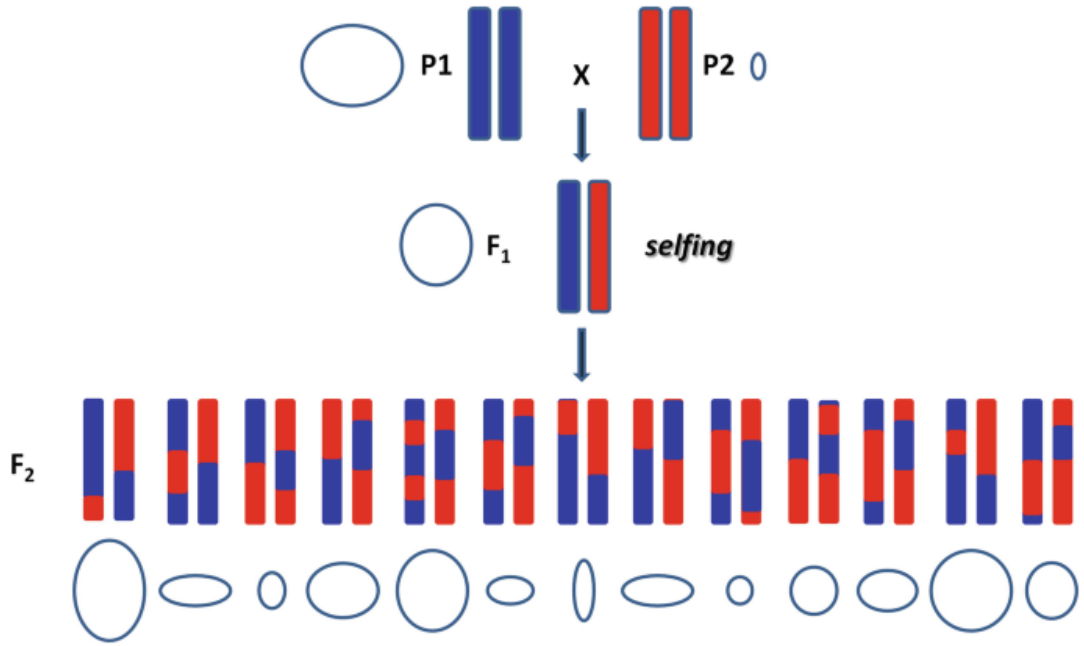
(F_2, DH, BC, RIL, NIL)

groups of individual plants with shuffled genomes from two parents.

2) genotyping and phenotyping of mapping population
3) construction of a linkage map
4) linkage analysis

- high statistical power to **detect 'major QTL'** and rare alleles that control a phenotype
- demands a limited number of markers and samples to be phenotyped

Linkage mapping

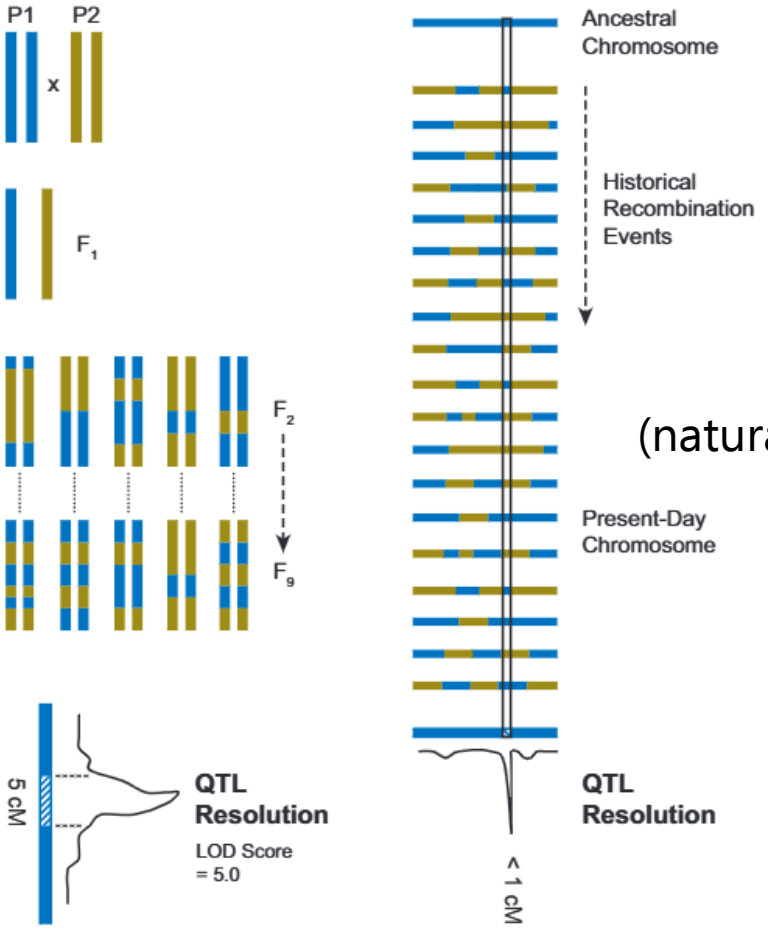


- experimental crosses, time consuming and laborious
- relatively narrow genetic base included, low mapping resolution
- long QTL intervals, 5-10 cM and contain many genes
- QTL specific to the bi-parental population and not useful in a wide genetic backgrounds

Strategies for QTL mapping

Linkage mapping

experimental populations
(bi-parental mapping populations)

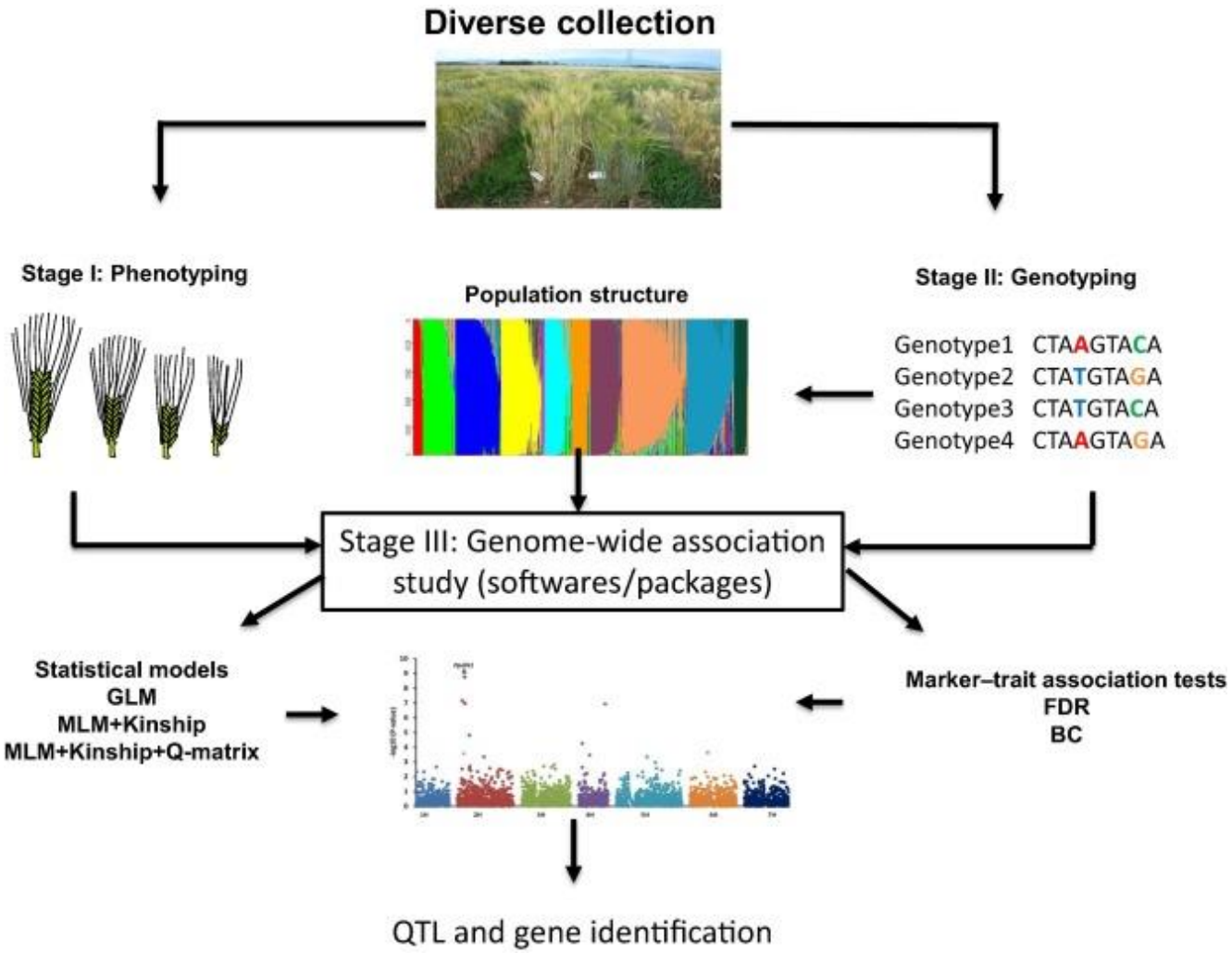


Association mapping

diverse lines of different origin
(natural populations or germplasm collections)

Soto-Cerda BJ, Cloutier S (2012)

Genome-Wide Association Studies (GWAS)



Alqudah et al. 2020

1. Define a population for analysis

2. Genotype the population

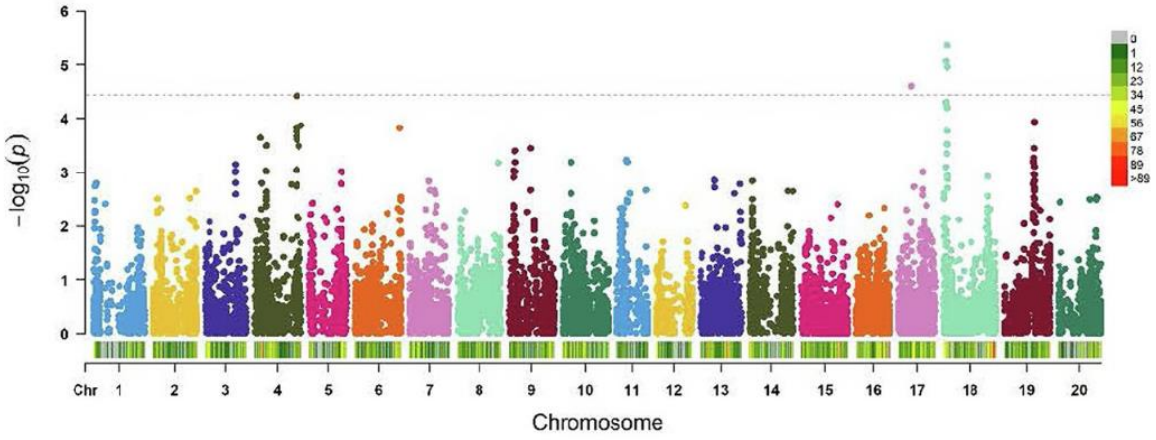
3. Population Structure/Relatedness

4. Phenotype the population

5. Statistical Analysis - correlation of phenotypic and genotypic data

Genome-Wide Association Studies (GWAS)

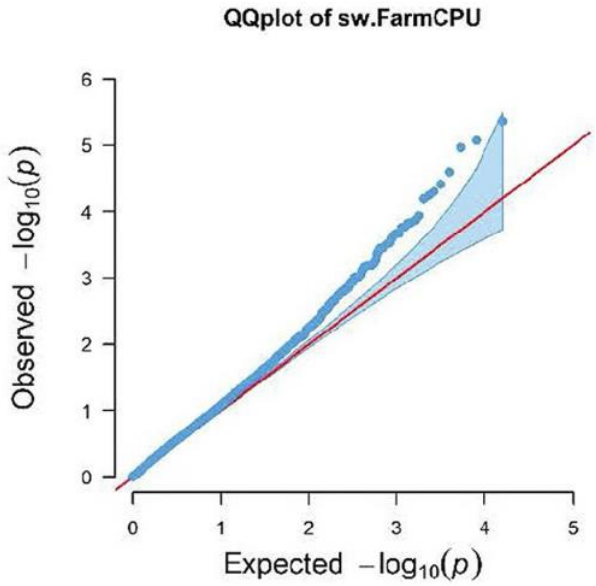
Manhattan (New York City) Skyline from New Jersey



Manhattan Plot

scatter plot display large number of data points (SNPs) and their significance in GWAS
(negative logarithm of the association p-value)

Q-Q plot (quantile-quantile plot)



Priyanatha et al. 2022 Front. Plant Sci. 13:866300

Soybean 100 seed weight GWAS analysis

Genome-Wide Association Studies (GWAS)

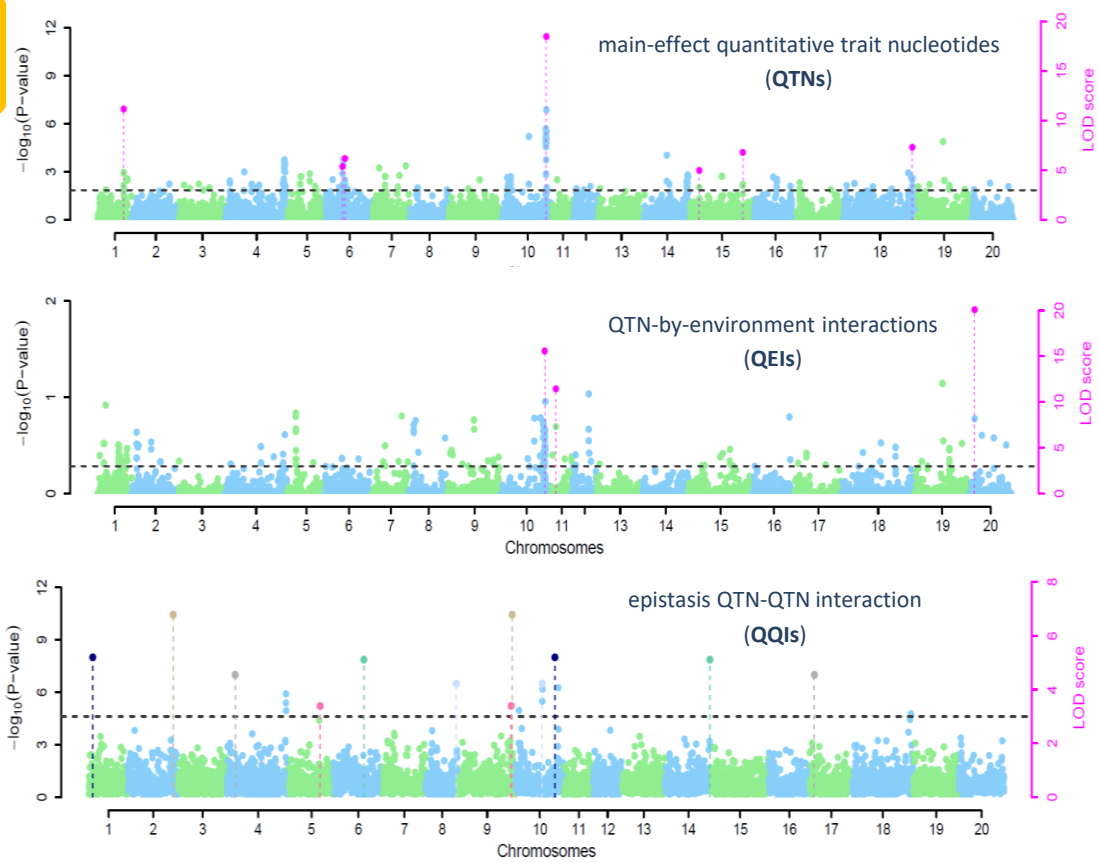
182 genotypes

Austria and Serbia
2020 and 2021

35k SNPs
(GBS, Wm82.a2)

3 Variance-component multi-locus random-SNP-effect Mixed Linear Model
IIIVmrMLM

Flowering time



Traits:

- **thousand seed weight** (TSW)
- **yield** (Y)
- **canopy cover** in vegetative (CCV) and reproductive phase (CCR)
- **grain quality** - protein content (P) and oil content (O)
- **plant architecture** - average internode length (AIL), node number (NN), plant height (PH), number of branches (NB)
- **local adaptation** - flowering time (P-R1), full maturity (P-R8), reproductive stage length (R1-R8)

Association (LD) mapping

(+) no need for the development of biparental populations

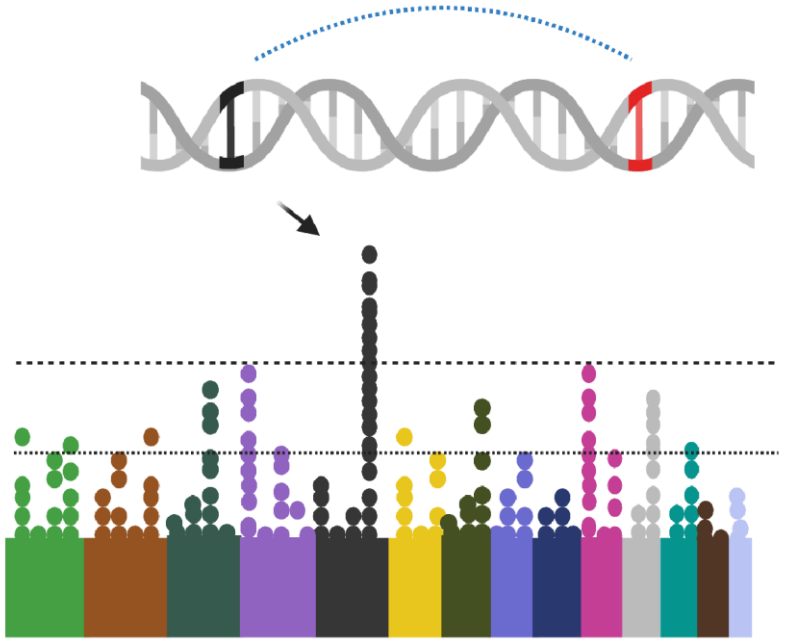
(+) high resolution mapping, take advantage of **LD** and **historical recombinations**

(+) availability of **broader genetic variations with wider background** for marker-trait correlation and **more QTLs underlying the traits**

(-) **less statistical power** (presence of alleles with low MAF, rare alleles)

(-) spurious signals of association derived from **population genetic structure** or non-random mating (**relatedness**)

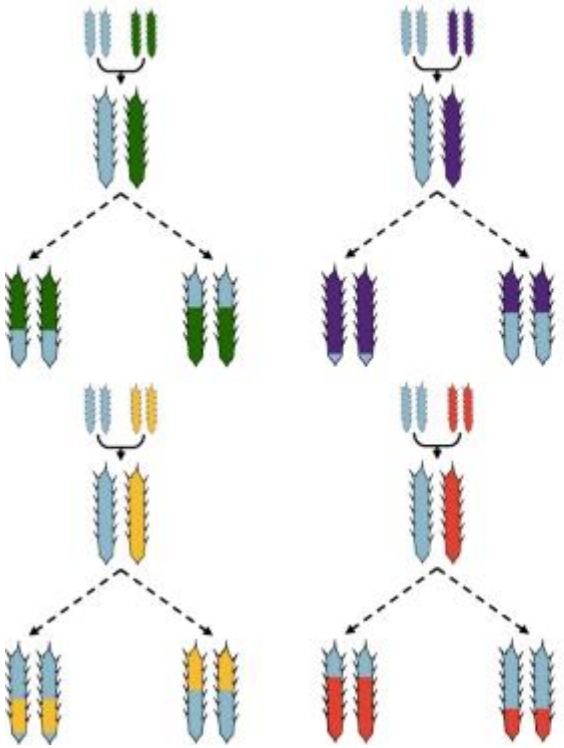
(-) large sample size – **significant phenotyping**



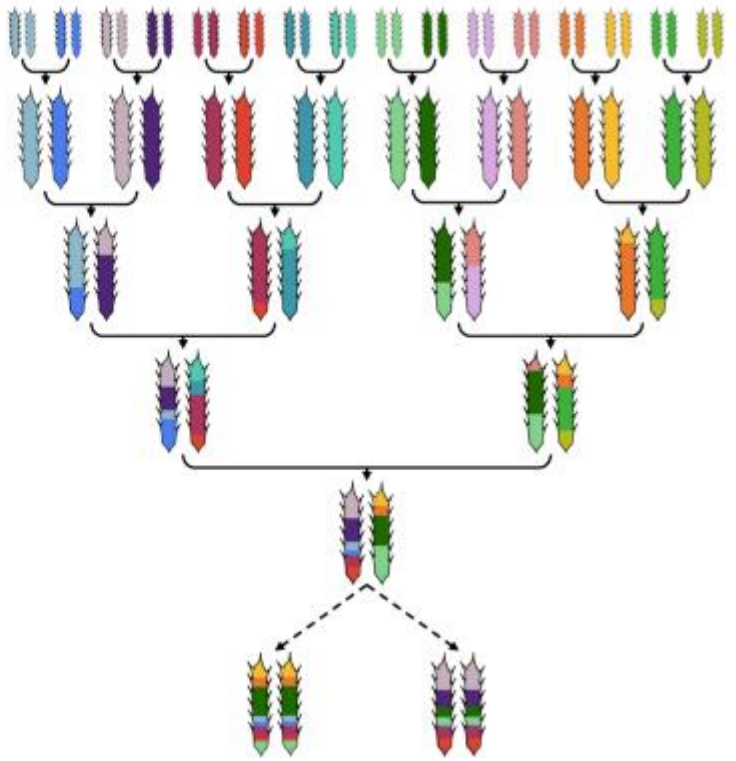
Advanced Mapping Populations

Feature	Combined Benefits for Breeding
Genetic Diversity	High diversity, access to a **wide range of alleles** from diverse backgrounds
Recombination	Increased recombination, improves **mapping resolution** and **gene discovery** power
Population Structure	Low structure, reduces **false positives** in QTL detection
Mapping Power	Enables **accurate and powerful QTL mapping** for both simple and complex traits
Allelic Richness	Helps in detecting **rare and common alleles** important for traits
Genomic Coverage	Supports development of **high-density genetic maps**
Final Product	Both provide **stable, permanent mapping populations** for repeated studies and developing superior crop varieties

Nested Association Mapping (NAM) panel

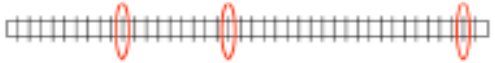


Multi-parent Advanced Generation Inter-Cross (MAGIC)



https://www.nature.com/articles/s41437-020-0336-6

Genomic selection



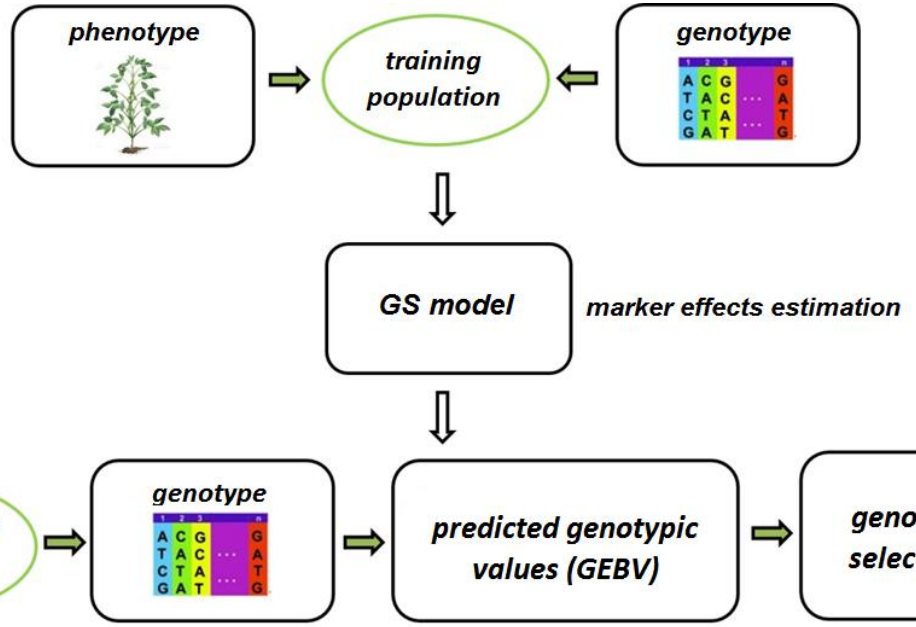
MAS

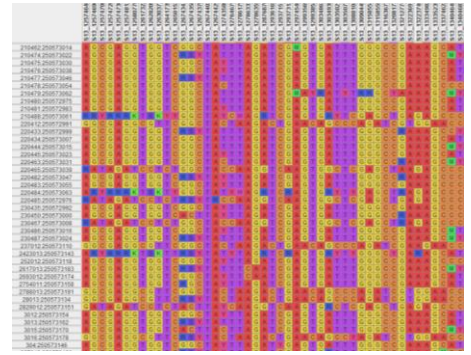
vs.

GS

large size of the mapped genomic regions
small contribution to phenotypic variation
genetic background effect

Genomic prediction is an approach that uses markers to predict the genetic value of complex traits in progeny for selection and breeding (Meuwissen et al. 2001)



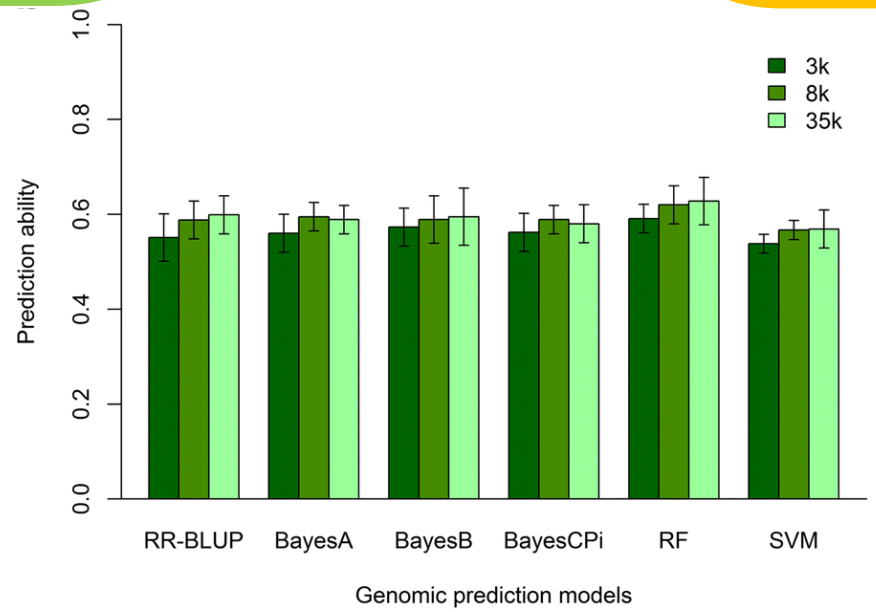


Genomic selection in soybean

227 breeding lines

3 years, seed yield

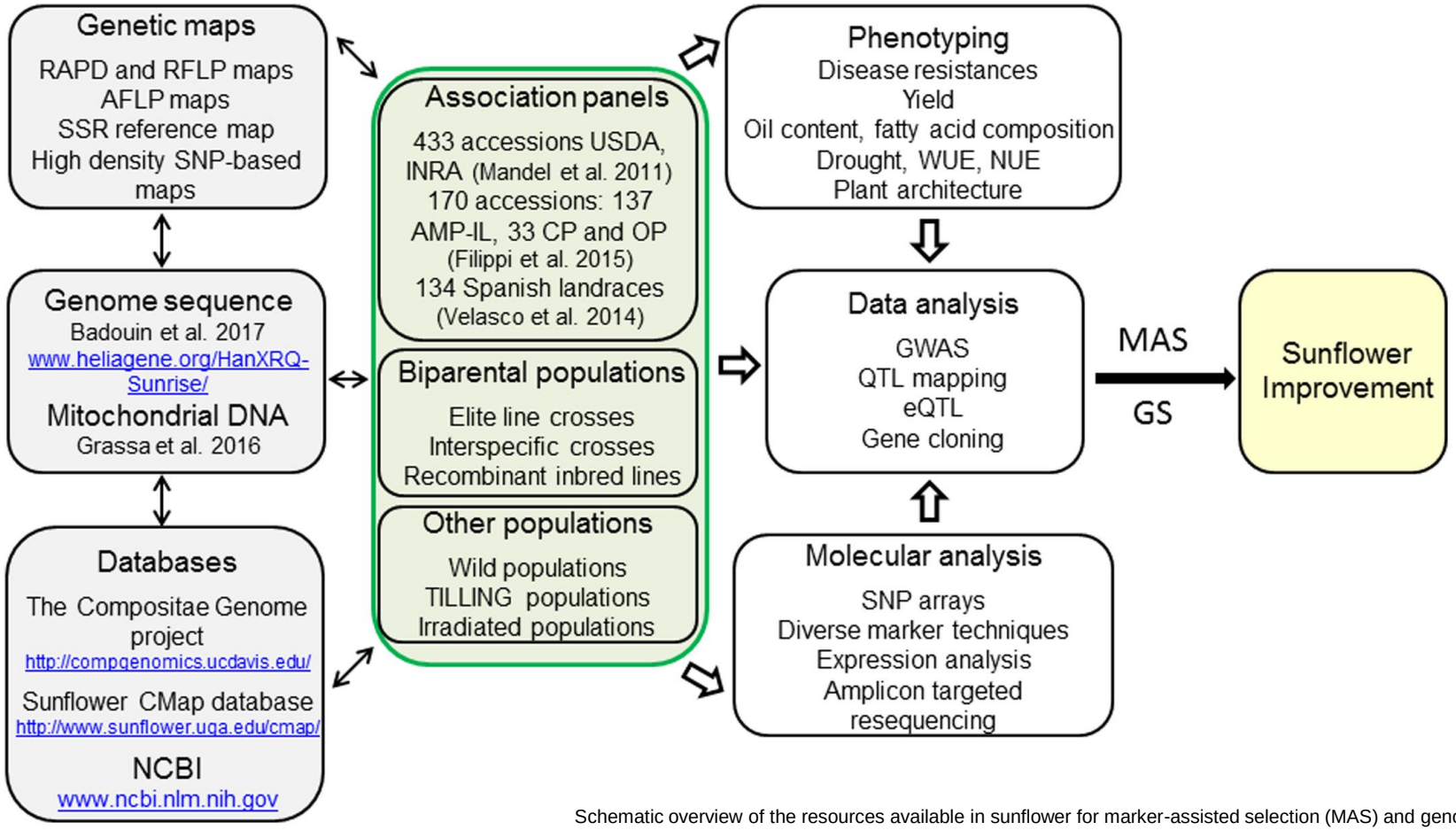
85k → 35k SNPs, GBS, Wm82.a2



Parametric models (*RR-BLUP, BayesA, BayesB, BayesCPi*) and non-parametric models (*RF, SVM*)

European prediction models

seed yield and protein content

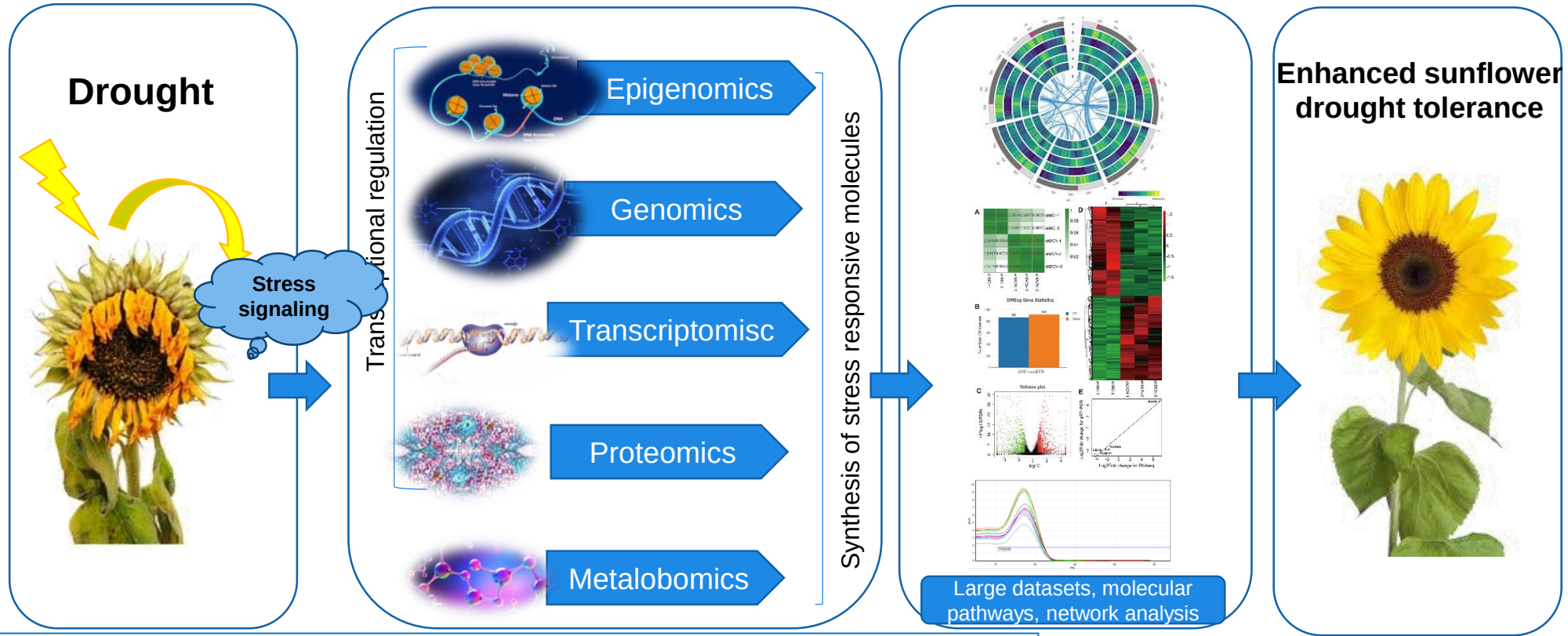


Schematic overview of the resources available in sunflower for marker-assisted selection (MAS) and genomic selection (GS). Diverse plant genetic resources for sunflower breeding are available representing a large genetic diversity that can be exploited for sunflower improvement. The access to the sunflower genome sequences, the large resources of SNP, being part of high resolution maps or SNP arrays, and the huge amount of expression data accelerate sunflower breeding by making the selection steps more efficient and precise.

Dimitrijevic, A., Horn, R., 2018. *Frontiers in Plant Science*, *8*, p.2238.

The Omics Revolution in Crop Breeding

Moving from studying a single word to analyzing an entire library to understand a story.



Omics refers to the study of an entire set of biological molecules, such as genes (genomics), RNA (transcriptomics), proteins (proteomics), or metabolites (metabolomics), within a biological system. It's a high-throughput, comprehensive approach to understanding the complete picture of these molecules and their inter-relationships, providing insights into the overall structure and function of a living organism at a particular level.

https://link.springer.com/chapter/10.1007/978-3-319-43033-1_1

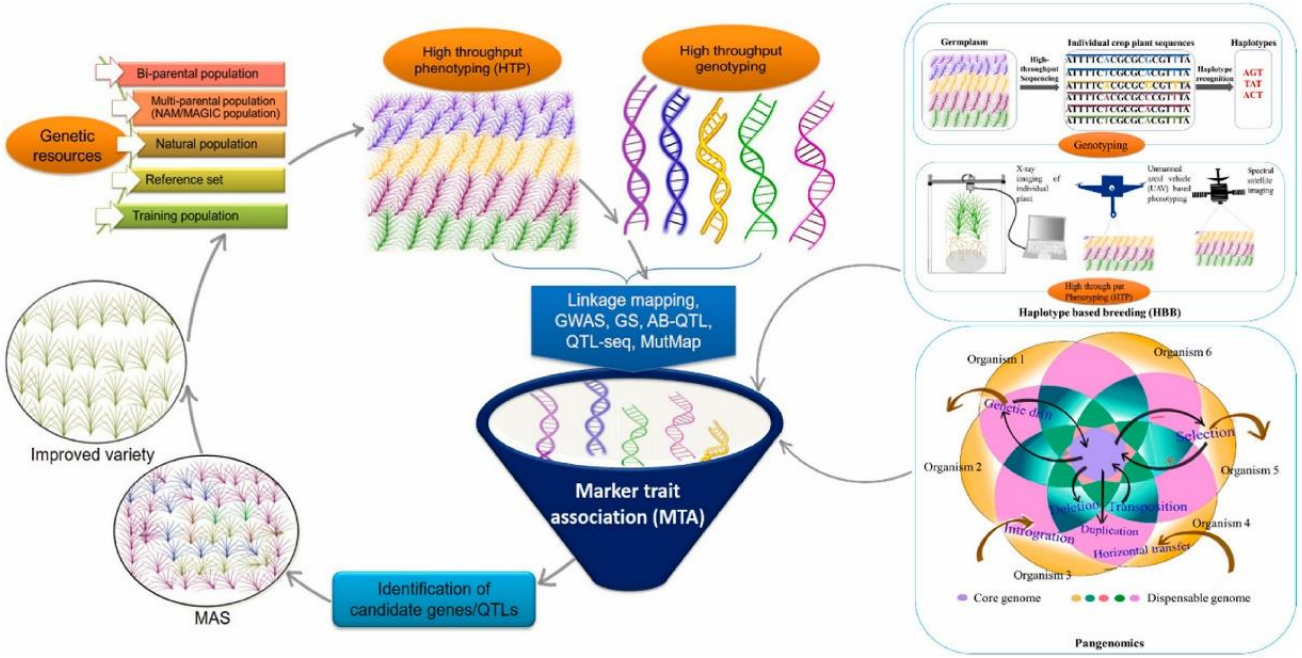
Genomics: The Complete Blueprint

Genomics is the study of an organism's entire set of genetic instructions, its **genome**.

Why for breeding? By analyzing a plant's entire genome, we can pinpoint genes and genetic variations (**polymorphisms**) associated with important traits like drought tolerance, disease resistance, and higher yield. It helps breeders understand the genetic potential of a plant before it's even grown.

Key Technologies: DNA Sequencing, which determines the precise order of nucleotides (A, T, C, G) in a DNA molecule. This generates massive amounts of data that can be used for:

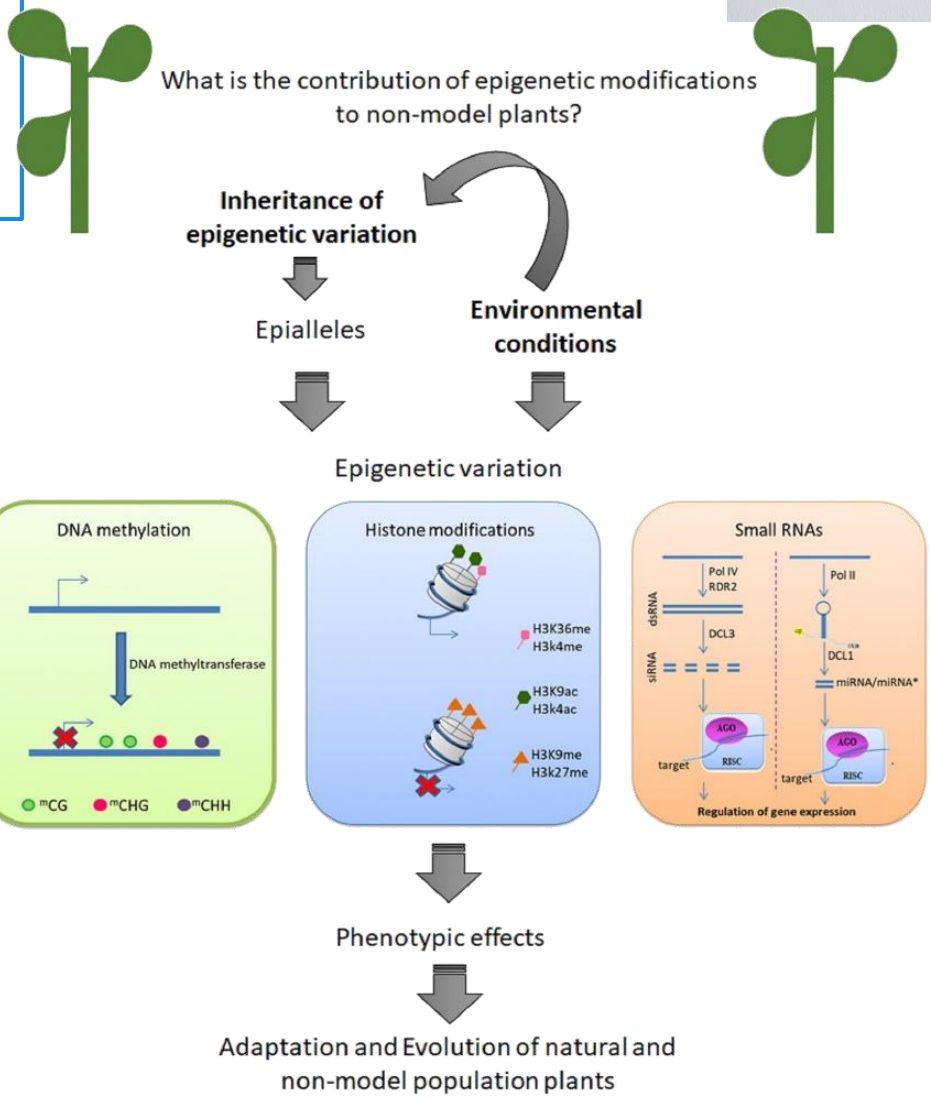
- **Genomic Selection:** Using DNA markers across the entire genome to predict the performance of a plant.
- **Genome-Wide Association Studies (GWAS):** A powerful method for linking specific genes to a trait by analyzing DNA samples from many different individuals.



https://www.cell.com/heliyon/fulltext/S2405-8440%2824%2911544-7

Epigenomics: Beyond the DNA Sequence

Epigenomics is the study of stable modifications to the genome and its associated proteins that influence gene activity. These modifications act as a layer of 'instructions' on top of the DNA blueprint, telling the cell which genes to use and which to ignore.



Key Components of the Epigenome:

DNA Methylation: This is the most studied epigenetic mark. It involves adding a methyl group (CH3) to a DNA base, typically cytosine. High levels of methylation in a gene's promoter region can act as a "switch," effectively silencing that gene.

Histone Modification: DNA is wrapped around proteins called **histones**. Chemical modifications to these histones (like acetylation or methylation) can either loosen or tighten the DNA, thereby "turning on" or "turning off" genes.

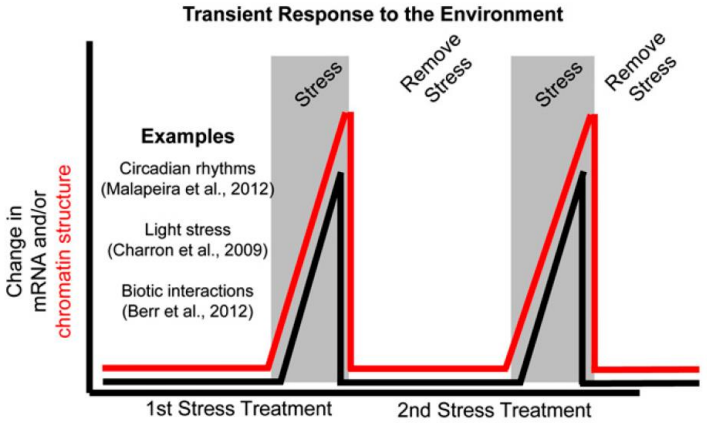
Non-Coding RNAs (ncRNAs): These are RNA molecules that are not translated into proteins. They play a crucial regulatory role by binding to specific DNA or messenger RNA sequences to control gene expression. Small ncRNAs, for example, can silence genes by degrading their transcripts or by promoting DNA methylation.

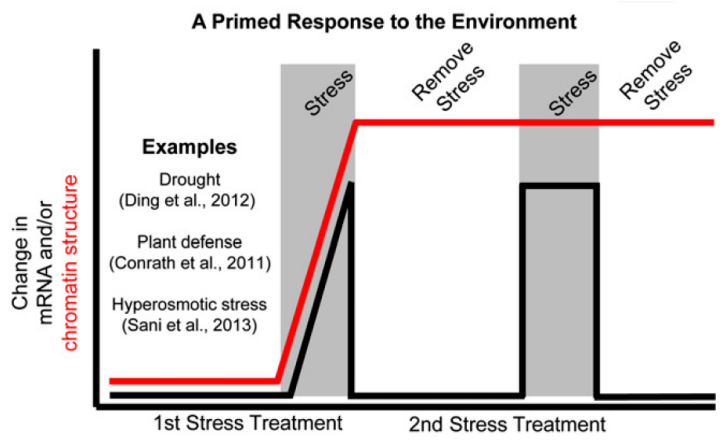
Thiebaut, F., et al. 2019 *Frontiers in plant science*, *10*, p.246.

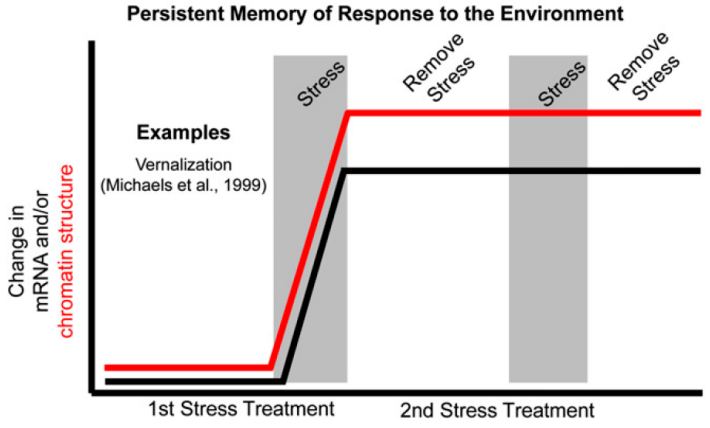
Epigenomics: Beyond the DNA Sequence

Why for Breeding?

Epigenetic changes are often influenced by the environment and can be passed down to offspring. This phenomenon, known as **transgenerational epigenetic inheritance**, means a plant's exposure to stress (like drought or heat) can result in an heritable adaptation, even though the DNA sequence has not changed. Understanding this can help breeders identify plants with an epigenetic "memory" of stress, leading to the development of more resilient crop varieties.







Key Technologies

A combination of advanced techniques is used to study the full epigenome:

- **Bisulfite Sequencing:** Used to map **DNA methylation** patterns across the genome.
- **Chromatin Immunoprecipitation (ChIP-seq):** Used to identify the locations of specific **histone modifications** and other DNA-binding proteins.
- **Small RNA Sequencing:** Used to identify and quantify the various **non-coding RNAs** involved in gene regulation.

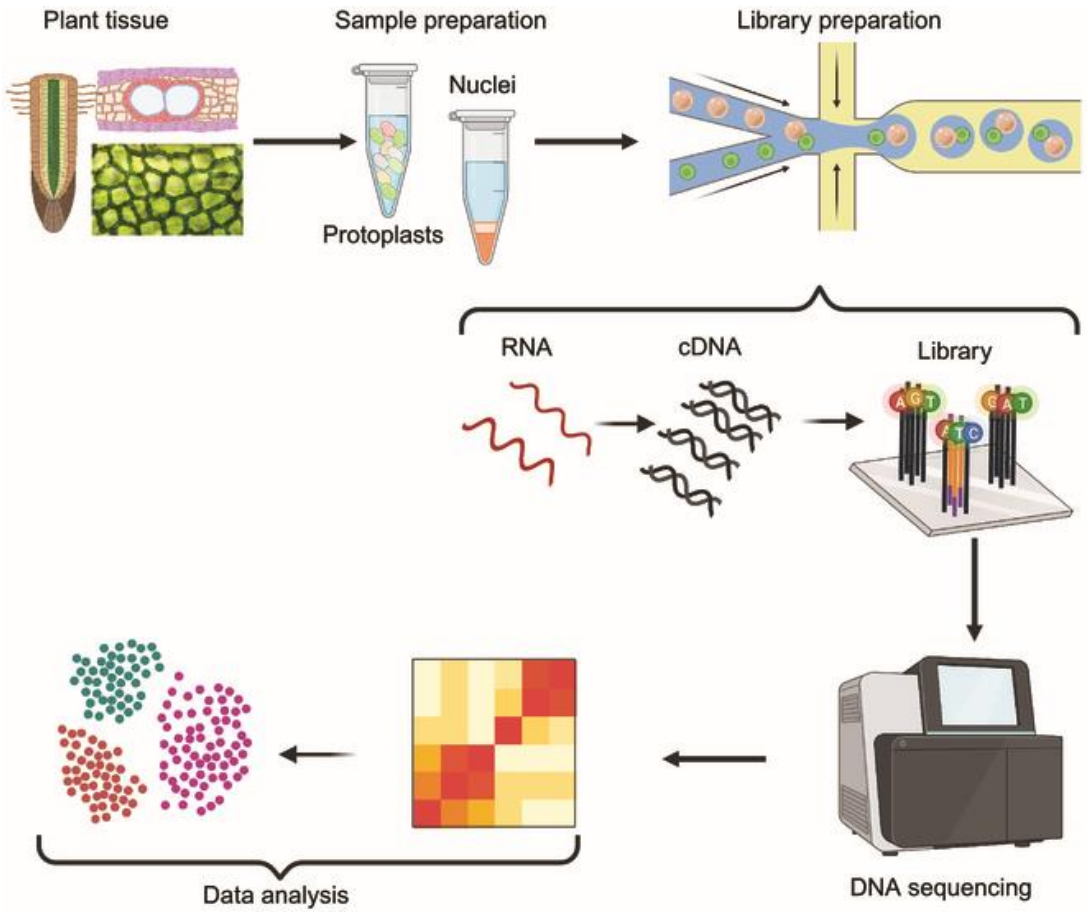
Eichten SR, et al. 2014. *Plant Physiol*. 165: 933-947.

Transcriptomics: The Active Genes

Transcriptomics is the study of all the **RNA transcripts** in a cell at a given time.

Why for Breeding?
By analyzing the transcriptome, breeders can see how a plant's genes are responding to its environment. This is particularly useful for studying how plants react to stress.

Key Technology
RNA Sequencing (RNA-seq): It involves converting RNA transcripts back into DNA and then sequencing them. This process provides a comprehensive snapshot of gene expression and can reveal the key genes and pathways that contribute to important traits like stress tolerance or nutrient efficiency.



https://spj.science.org/doi/10.34133/bdr.0029

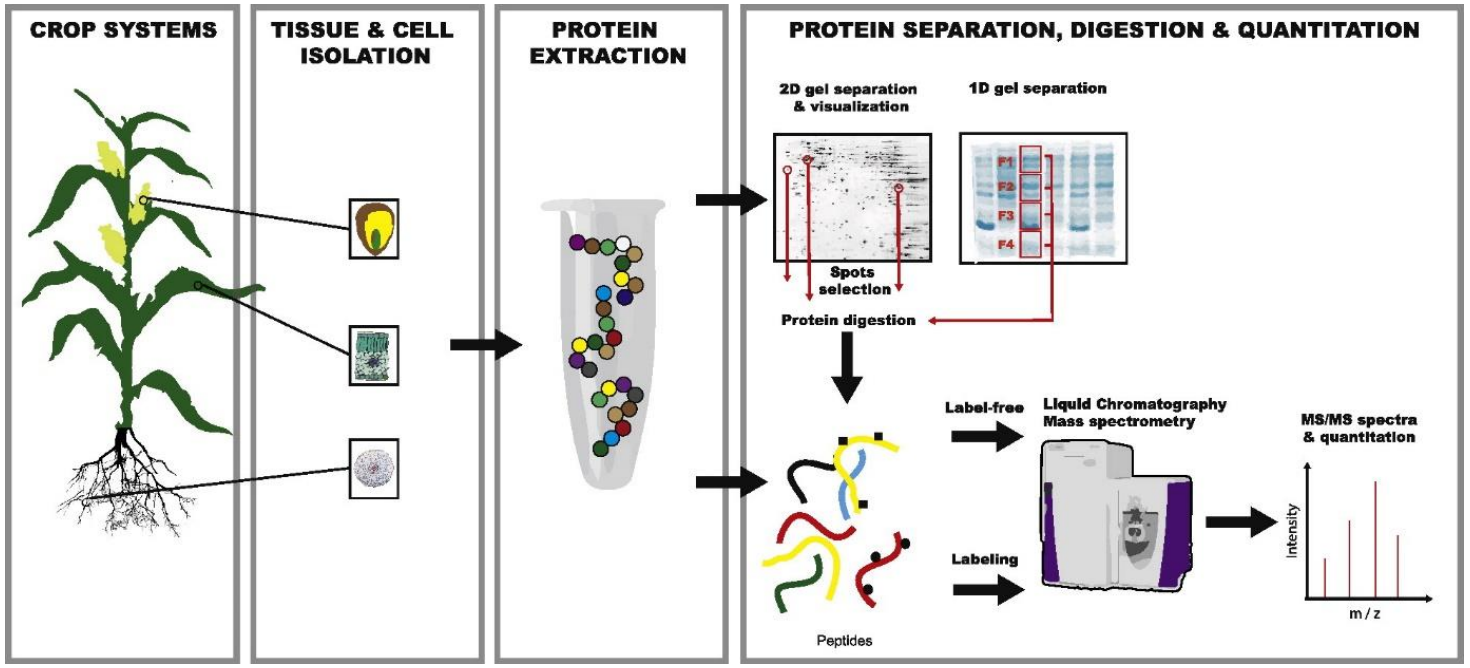
Proteomics: The Protein 'Workforce'

Proteomics is the large-scale study of **proteins**.

Proteins are the **"workhorses"** of the cell. They are enzymes that build and break down molecules, they are structural components, and they are signals that communicate between cells.

Why for breeding? By analyzing which proteins are present and in what amounts, we can directly see the plant's physiological response to stress.

Key Technology: The primary tool for proteomics is **Mass Spectrometry**, which identifies proteins based on their mass and charge



https://www.sciencedirect.com/science/article/abs/pii/S1874391913003126

Metabolomics: The Final Result

Metabolomics is the large-scale study of all the small-molecule metabolites in a plant. Metabolites are the final products of cellular processes, such as sugars, amino acids, lipids, and vitamins. They are the most direct reflection of a plant's physiological state and its interaction with the environment.

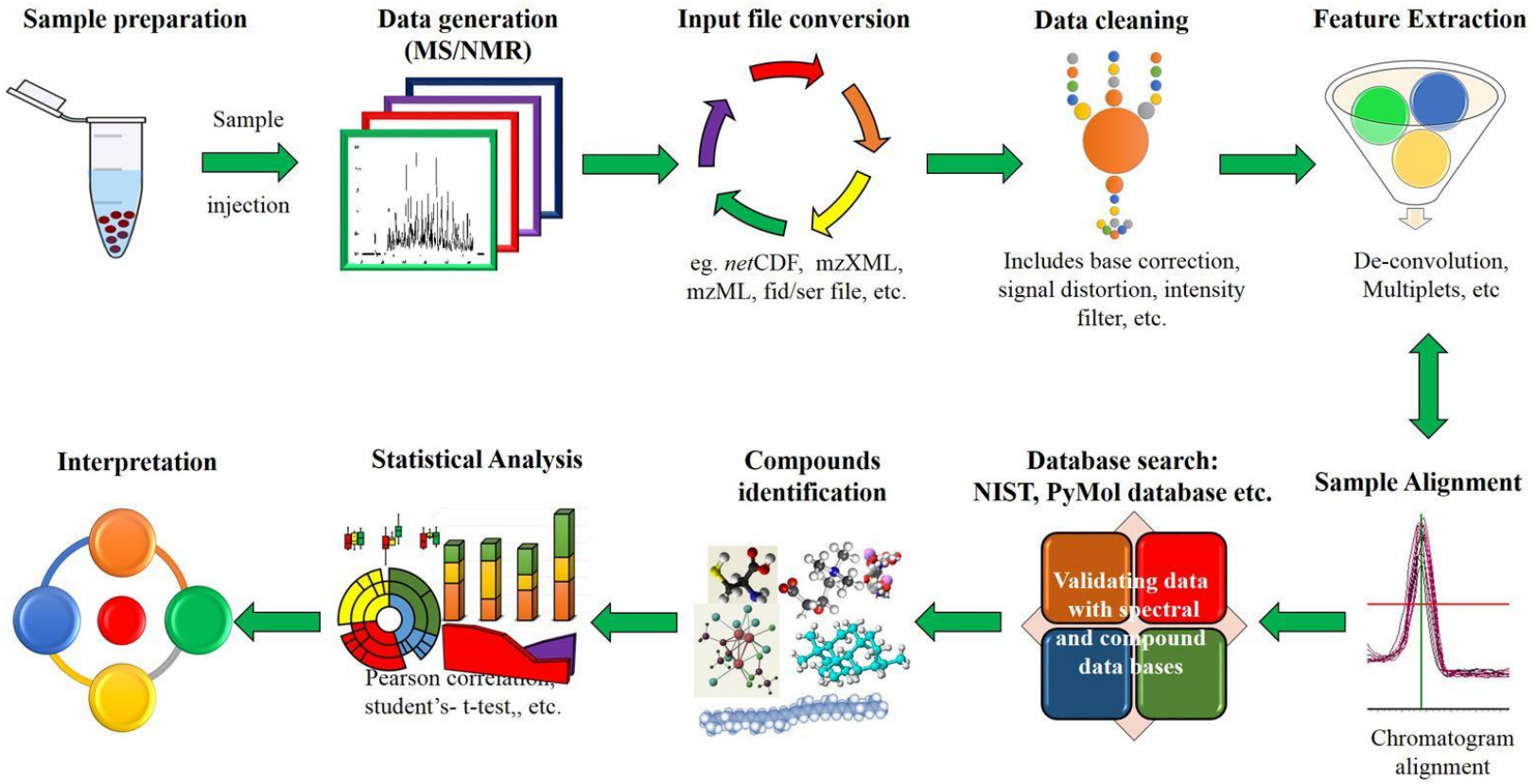
Why for Breeding?

- Metabolomics provides a direct window into the plant's health, quality, and stress responses.
- By analyzing a plant's metabolic profile, breeders can identify key metabolites associated with desirable traits like flavor, nutritional content, or resistance to a pest.

Key Technologies

•**Mass Spectrometry (MS)**: This is the main tool used to separate and identify metabolites based on their mass-to-charge ratio. It provides a highly detailed "fingerprint" of the plant's metabolic state.

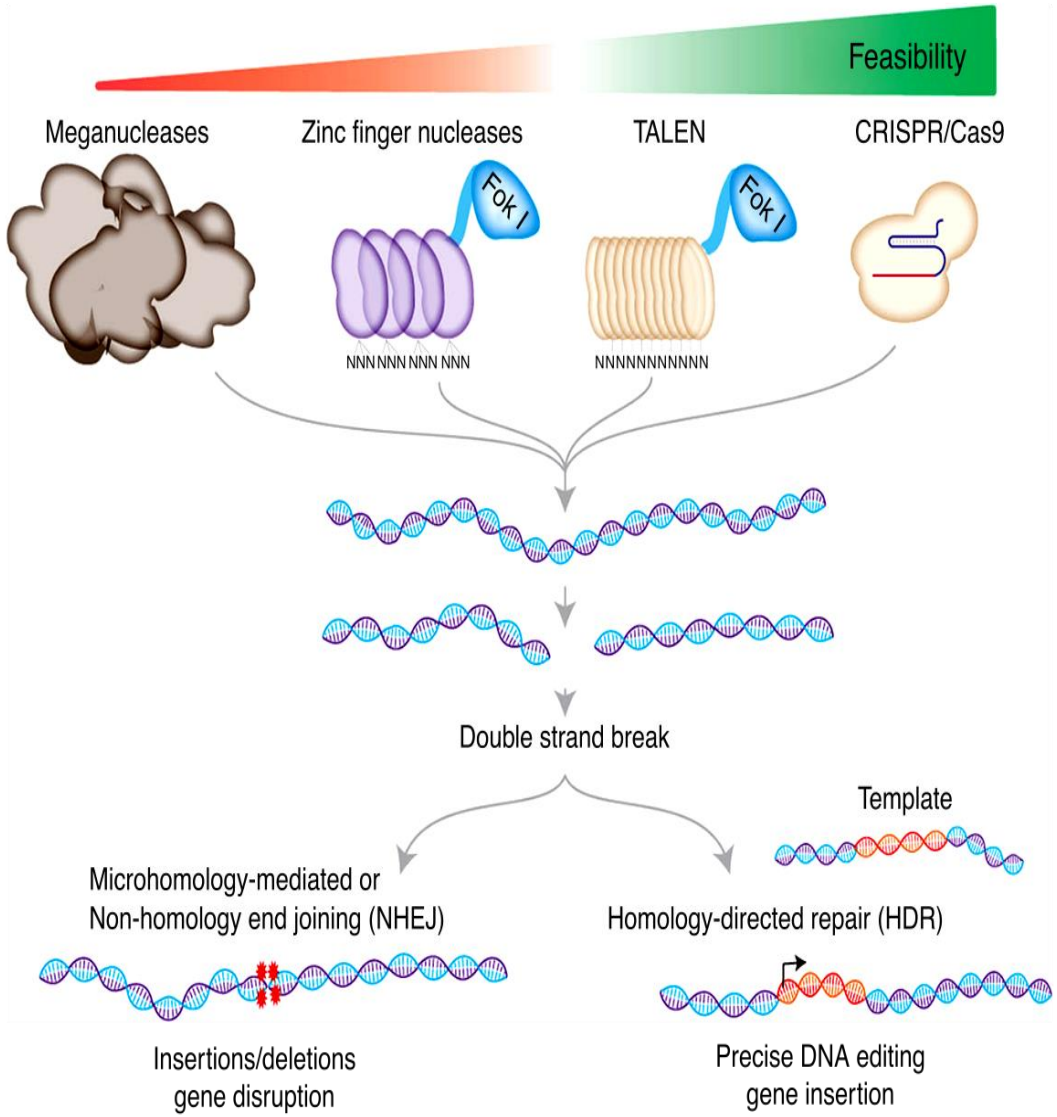
•**Nuclear Magnetic Resonance (NMR)** Spectroscopy: This technique uses magnetic fields to identify and quantify metabolites, providing a non-destructive way to analyze samples.



https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2017.01302/full

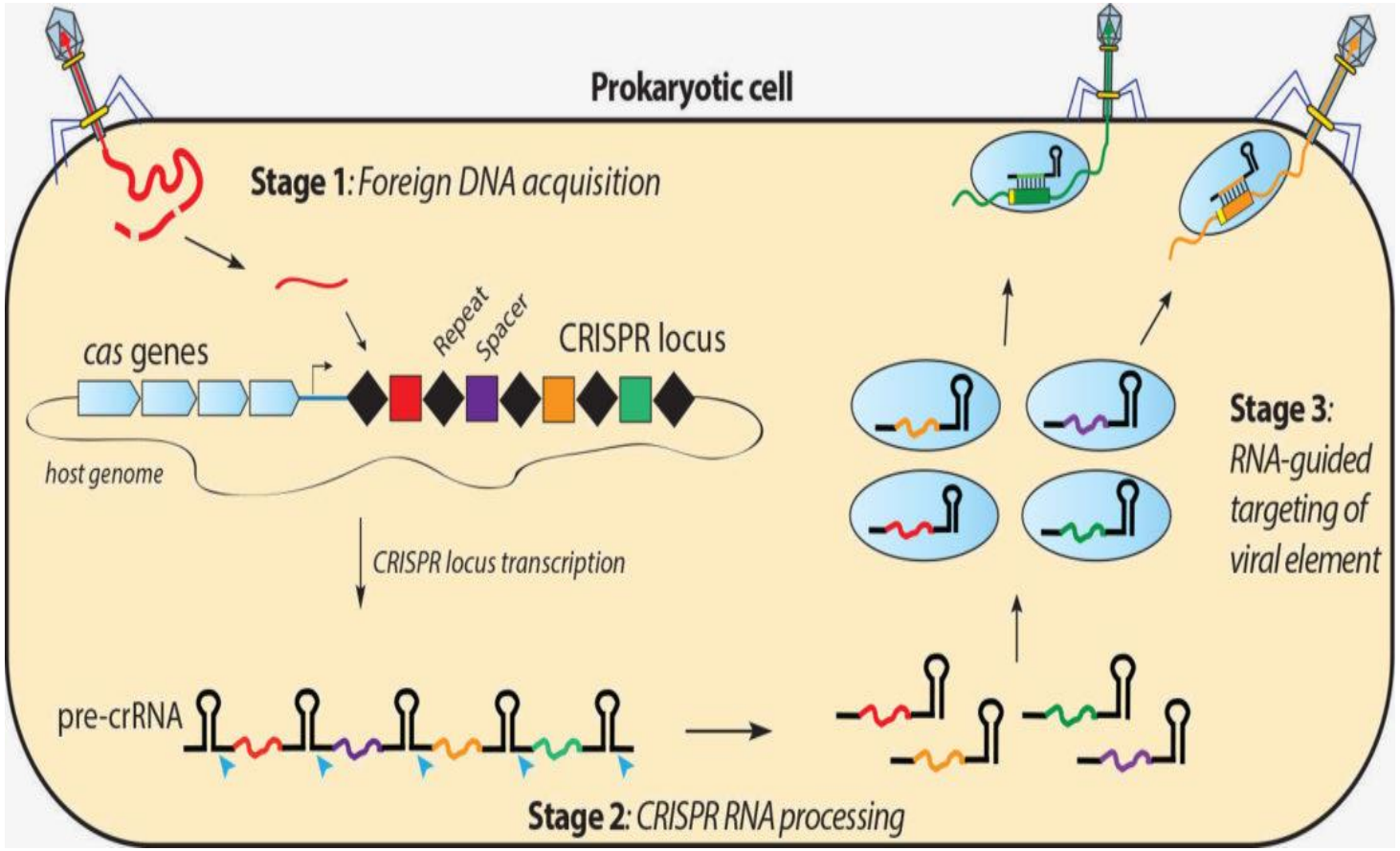
Genome editing

Gene editing allows scientists to make precise, targeted changes to a plant's DNA sequence. It is a powerful tool for crop breeding, enabling the introduction of desirable traits with high accuracy.
The key to these methods is creating a **double-strand break** at a specific location in the genome



CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

- A groundbreaking gene editing system that is far more efficient and user-friendly than previous methods.
- It consists of two key components: a **guide RNA (gRNA)** and a **Cas9 enzyme**.

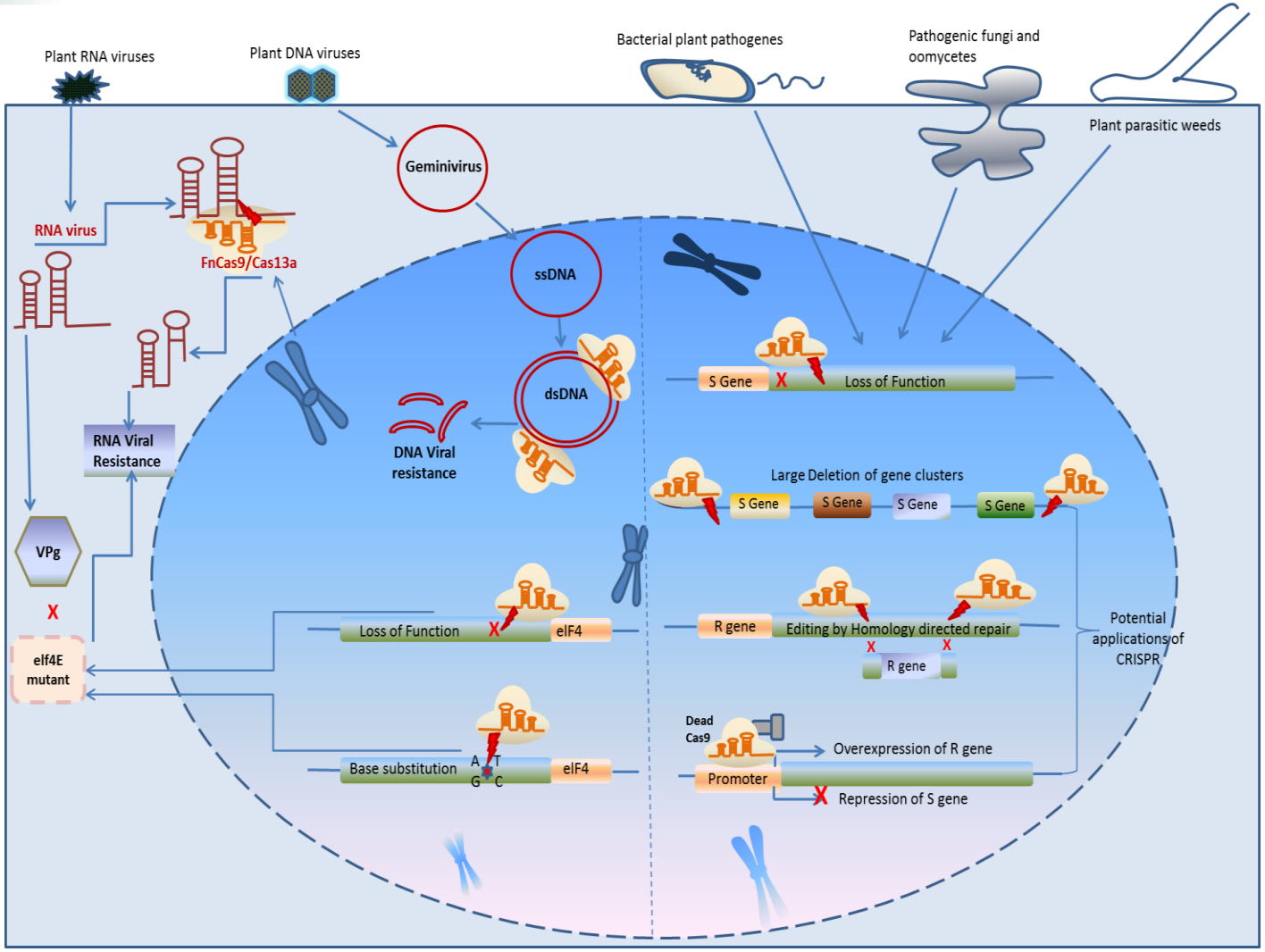


How it works:

•The **gRNA** is a custom-designed RNA molecule that acts like a GPS, guiding the system to a precise location on the DNA sequence. It is a simple, single molecule that can be easily changed to target any gene.

•The **Cas9 enzyme** is the "molecular scissors" that creates a double-strand break at the target site identified by the gRNA.

Versatility: CRISPR can be used for a wide range of applications, including gene knockouts, gene insertions...



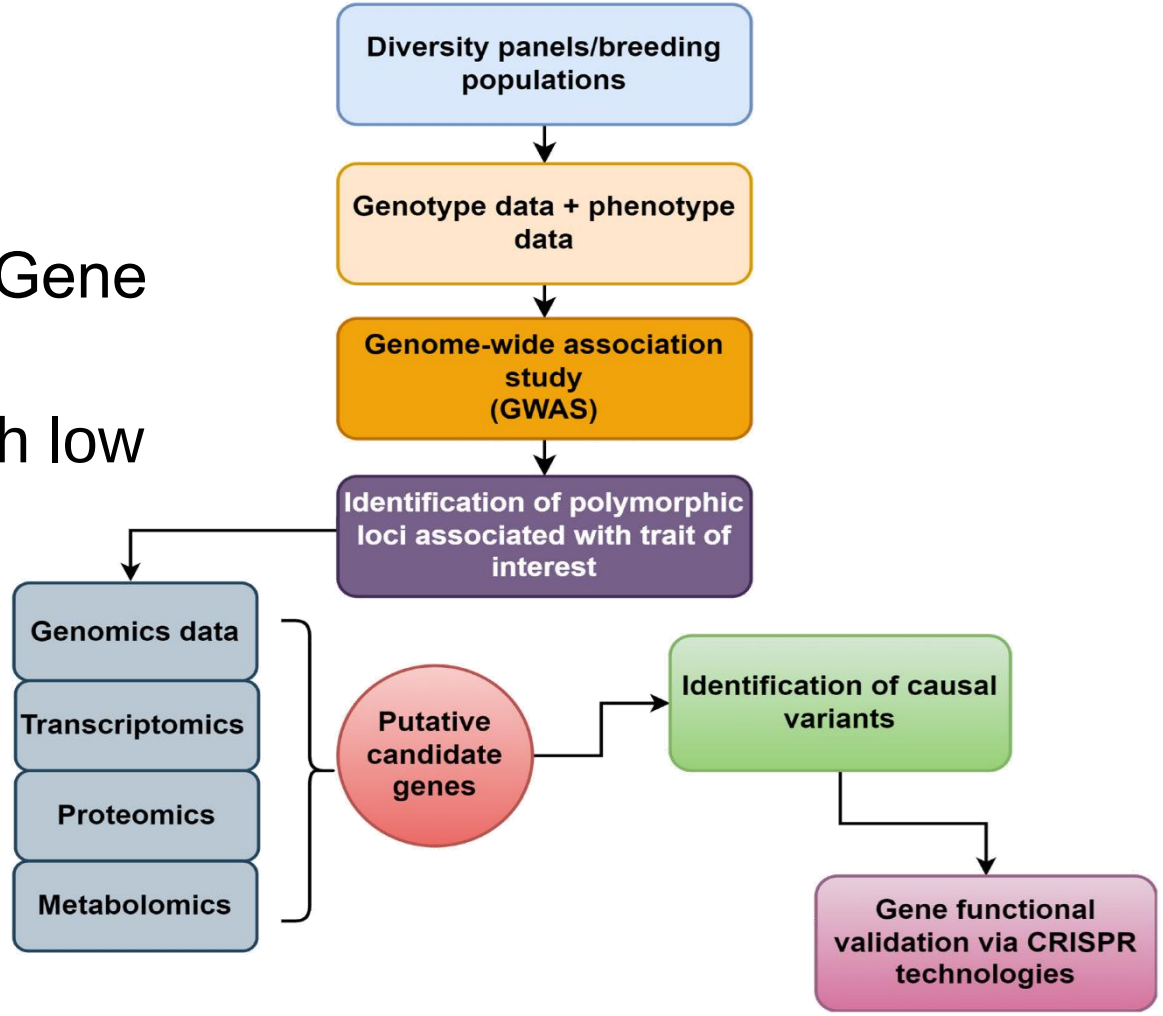
https://www.youtube.com/watch?v=4YKFw2KZA5o

Which strategy should I use for my breeding?

One gene (MAS, or MABC)

Multiple genes with high heritability ≤ 20 (MA Gene Pyramiding, MA Recurrent Selection)

Quantitative traits controlled by polygenes with low heritability > 20 (Genomic Selection)



Thank you for your attention!

Any questions?